IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.
Cite as T-ASE paper.

Naturalistic Robot-to-Human Bimanual Handover in Complex Environments Through Multi-Sensor Fusion

Salih Ertug Ovur, *Graduate Student Member, IEEE*, and Yiannis Demiris, *Senior Member, IEEE*

***Abstract*— Robot-human object handover has been extensively studied in recent years for a wide range of applications. However, it is still far from being as natural as human-human handovers, largely due to the robots' limited sensing capabilities. Previous approaches in the literature typically simplify the handover scenarios, including one or more of (a) conducting handovers at fixed locations, (b) not adapting to human preferences, or (c) only focusing on single-arm handover with small objects due to the sensor occlusions caused by large objects. To advance the state of the art toward a human-human level of handover fluency, this paper investigates a bimanual handover scenario in a naturalistic, complex setup. Specifically, we target robot-to-human box transfer while the human partner is on a ladder, and ensure that the object is adaptively delivered based on human preferences. To address the occlusion problem that arises in a complex environment, we develop an onboard multi-sensor perception system for the bimanual robot, introduce a measurement confidence estimation technique, and propose an occlusion-resilient multi-sensor fusion technique by positioning visual perception sensors in distinct locations on the robot with different fields of view. In addition, we establish a Cartesian space controller with a quaternion approach and a leader-follower control structure for compliant motion. Four distinct experiments are conducted, covering different human preferences (such as the box delivered above or below the hands) and significant handover location changes once the process has begun. For validation, the proposed multi-sensor fusion technique was compared to a single-sensor approach for both top and bottom sensors separately, and to simple averaging of both sensors. 30 repetitions were performed for each experiment (four experiments, four methods), the equivalent of 480 handover repetitions in total. Multi-sensor fusion approach achieved a handover success rate above 86.7% for all experiments by successfully combining the strengths of both fields of view for human pose tracking under significant occlusions without sacrificing handover duration. In contrast, due to the occlusions, the single-sensor and simple averaging approaches completely failed during challenging experiments, illustrating the importance of multi-sensor fusion in complex handover scenarios.**

Manuscript received 2 March 2023; accepted 27 April 2023. This article was recommended for publication by Associate Editor L. Liu and Editor J. Yi upon evaluation of the reviewers' comments. This work was supported in part by the Imperial College London President's Ph.D. Scholarship, U.K. Research and Innovation (UKRI), under Grant EP/V026682/1; and in part by the Royal Academy of Engineering Chair in Emerging Technologies. *(Corresponding author: Salih Ertug Ovur.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

The authors are with the Personal Robotics Laboratory, Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K. (e-mail: e.ovur21@imperial.ac.uk; y.demiris@imperial.ac.uk).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TASE.2023.3284668.

Digital Object Identifier 10.1109/TASE.2023.3284668

***Note to Practitioners*—This paper is motivated by enabling naturalistic robot-to-human bimanual object handovers in complex environments, which is a challenging problem due to occlusions. Existing approaches in the literature do not benefit from multi-sensor fusion to handle occlusions, which is essential in such physical human-robot interaction scenarios. To this aim, we have developed a multi-sensor fusion technique to improve the perception capabilities of robots with respect to human co-workers. The developed framework has been tested with Microsoft Azure Kinect sensors and a bimanual mobile Baxter robot, but it can be adapted to any depth perception sensor and bimanual robotic platform. Furthermore, the introduced multi-sensor fusion technique is comprehensive and generic, as it can be applied to any intermittent sensor data, such as human pose tracking via RGBD sensors. The presented approach shows that increasing the field of view of robots' perception used with enhanced data fusion could drastically improve the robot's sensing capability. For future work, data fusion can be improved by introducing Bayesian filters, and the system can be validated with different sensors and robotic platforms. Moreover, the handover detection method of physical interaction could further benefit from the incorporation of force sensors.**

***Index Terms*— Human-robot interaction, human-robot object handover, multi-sensor fusion, occlusions.**

I. Introduction

HUMAN-ROBOT interaction (HRI) scenarios such as robot-human object handover have become increasingly popular as the prevalence of collaborative robots has expanded in recent years [1], [2], [3]. Although the earliest applications in the field of human-robot object handover date back to the 1990s [4], [5], and numerous applications have been developed so far, there is still significant research that is needed to reach the naturalness of human-to-human object handovers. State-of-the-art methods mostly focus on simple scenarios, including single-arm handover with small objects, fixed handover locations, and not always perceiving and adapting to human preferences. One of the primary reasons for these simplifications is that robots lack the same complex sensory systems and awareness as humans have [6] for perceiving the body state and preferences of their human co-workers. This is especially noticeable when transferring more complicated objects in complex environments, where full or partial occlusions are present due to the close interaction distance, large-sized objects, ladders, robotic manipulators, and challenging camera view angles [7].

Studies in this field can be grouped into two main categories: a) the prehandover phase, which includes communication, grasp planning, perception, handover location,

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

and motion planning and control stages, and b) the physical handover phase, which includes grip force modulation and error management [8].

The prehandover phase starts with the communication stage to begin a collaborative action and continues to coordinate the handover after it has begun [9]. The communication stage is followed by understanding grasping choices and grasp planning [10], including studies about hand placement on objects during the handover [11]. Further research is done on grasp generation with a proposed human-to-robot unknown object handover by utilizing the developed hand, and object segmentation [12]. Moreover, [13] developed an object-independent human-to-robot handover framework to deliver previously unknown objects by using the modified generative grasping convolutional neural network.

Regarding the handover location, many studies have already examined fixed handover locations [14], [15], [16]. However, only a few of these studies considered the problems involved in online handover locations. According to [8], one of the crucial open challenges in human-robot object handover is that the handover location should be adjusted online. The location of the hand is employed in [17] to execute online handover location in various natural situations. Furthermore, recent studies showed that object release behaviors, such as proactive release, a form of online handover location, result in shorter completion times [18]. In the direction of learning the optimum online handover location for robot-human frameworks, human-human demonstrations are also considered [19]. In [20], a multitask variational autoencoder jointly forecasts the poses of the human participants in a handover as well as the orientation of the object as held by the giver during the handover. Despite the fact that these studies addressed the challenge of online handover location estimation, their approach is confined to single-arm handovers and is not resistant to occlusions caused by large objects during the handover.

Failure handling and grip force regulations are explored as part of the second phase of the handover, the physical handover phase. To prevent the item from falling, some researchers have used object acceleration as a warning of approaching handover failure, and have suggested a suitable re-grasping mechanism [21]. Another technique considers the tactile data from the robot hand grippers [22]. Moreover, grip force regulations are studied for successful and smooth handover delivery [3].

Beyond the object handover phases, researchers have recently begun to focus on more natural handover applications contrary to an idealistic laboratory environment, such as in human-human handover scenarios. Three natural environments were studied in [17], such as 'Lying under the car', which simulates lying on one's back and working, 'Engine Bay', which simulates working on the engine of a car with the person bent over, and 'Hydraulic Lift' which simulates working under a car on a hydraulic lift. Another realistic setup is investigated in [23], where object handover is done while a human worker conducting maintenance tasks stands on a ladder. In [24], Gaussian process regression is used for latent reward estimation of the absolute and preference

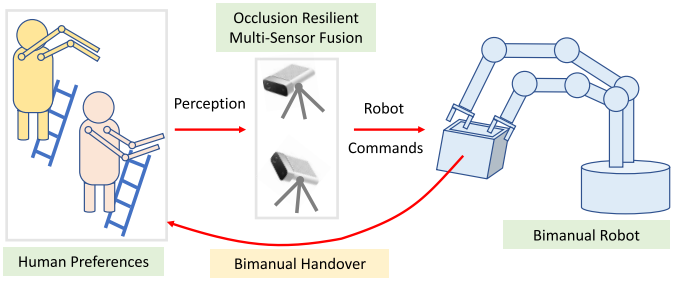
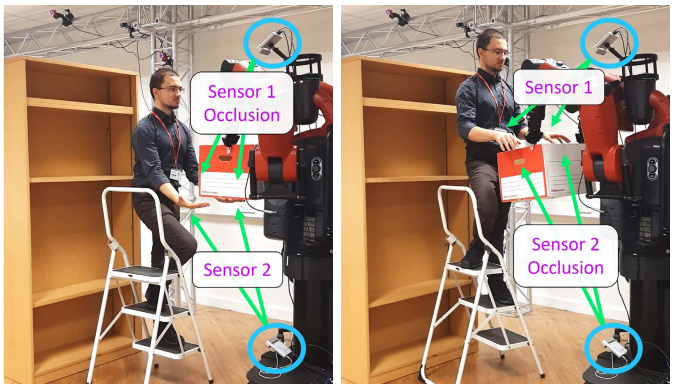


Fig. 1. Overview of the proposed technique for robot-to-human bimanual handover in complex environments, considering human preferences and enduring occlusions via the proposed multi-sensor fusion approach.



(a) (b)

Fig. 2. Different human preferences illustration for bimanual object handover with a large box from (a) above or (b) below the hands. Common approaches that rely on a single sensor cannot easily compensate for these diverse preferences due to the occlusions visible in both scenarios. As a solution, we present a multi-sensor fusion with distinct fields of view for achieving occlusion-resilient dynamic bimanual handover.

human feedback, and integrates reward learning with contextual policy search in order to learn dynamic robot-to-human object handover. However, only a limited number of studies focus on the challenges associated with bimanual handovers.

According to the literature, there is still a significant gap in solving bimanual object handover, especially when considering realistic scenarios in complex environments demanding online handover location determination [8]. In this paper, we developed a robust multi-sensor system (Fig. 1) that can work around occlusions. We did this by taking advantage of the strengths of the different fields of view perception sensors and combining pose tracking data with the proposed algorithms. Furthermore, we utilized these algorithms for performing robot-to-human bimanual object handover towards the ultimate goal of achieving naturalistic human-robot interaction. In order to achieve this goal, we developed a pipeline where the human pose is extracted and fused from two depth sensors, and the hands' pose is used to update the dynamic handover location. Despite the fact that occlusions frequently occur in such large object handover cases and complex environments, the proposed data fusion algorithm was able to track hands continuously, as illustrated in Fig. 2. Furthermore, hand orientation is used to deliver objects in accordance with human preferences. For validating the proposed fusion method,

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

we have used a robot-to-human bimanual object handover application where the human operator receives a large-size box and stands on a ladder, demonstrating a challenging environment where the robot's assistance would be genuinely useful. In order to assess the importance of an occlusion-resilient framework for large object delivery, online handover location, and adaptive framework to human preferences, four sets of experiments are performed with the four different techniques, including the proposed one and other compared methods. The results demonstrate the advantages of the proposed technique, which assigns weights to the sensor data based on the estimated measurement confidence, over conventional techniques that rely on a single sensor approach or simple averaging of the sensor data, which are prone to failure in the presence of occlusions. The contributions of this paper include:

- A bimanual robot-to-human object handover technique is proposed for handing over big objects in naturalistic environments. Moreover, human preferences are used to adapt handover strategy to improve the human experience.
- Multi-sensor perception system is designed to improve the perception capabilities of the robot. Accordingly, a measurement confidence estimation algorithm is proposed. These confidences are used to enhance the weighted averaging data fusion algorithms, making them resilient to occlusions that are inevitably occurring due to the large object handover and complex environments.
- An online handover location update framework is proposed using the introduced occlusion-resilient multi-sensor fusion module. This framework allows the human operator to alter the handover position during the action, enabling him/her to take over the object more comfortably. Additionally, this approach provides adaptability to human operator demands.

II. Related Works

A. Multi-Sensor Fusion

In recent decades, the topic of multiple sensor fusion has garnered significant research interest as it has been discovered that the combination of information from multiple sensors can provide more precise and accurate information [25], [26], [27]. Practical applications of the multi-sensor approach are now spreading among the different domains. One of the most common applications in this field contains multi-camera based applications which are especially encountered in the field of surveillance for tracking people [28]. In order to tackle occlusions and lack of visibility, multi-view information acquired from the multi-cameras is used to track people consistently [29]. Furthermore, surveillance techniques are becoming increasingly prevalent in our daily lives, such as recently introduced surveillance-powered, no-checkout cashierless stores [30]. Modern robots have also begun to benefit from multiple camera settings. For example, quadruped robot SPOT[1] utilizes five stereo cameras with a 360-degree field of view, to navigate and avoid obstacles effectively.

From an academic perspective, new fusion techniques are constantly being researched. Multiple sensor fusion through Kalman filters is applied to five Microsoft Kinect[2] sensors and results are compared to the ground truth obtained from the motion capture system in [31]. Similarly, multiple Kinect sensor data are fused for dance analysis by using a hidden-state conditional random fields classifier in [32]. Another study to handle the occlusion problem or noisy estimations was conducted by [33], via estimating measurement confidence and then using these estimations with fuzzy logic system integration as an input to the energy function for the final estimation. Reference [34] also positioned four Kinect sensors to the corners and used an information-weighted consensus filter combined with a multiple interactive model for retrieving reliable human pose estimation. In our previous works, we used multiple Leap Motion Controllers[3] (LMC) with adaptive Kalman filters to overcome occlusion problems [35]. In other previous research, we have utilized multiple LMCs to achieve reliable gesture recognition for enhancing HRC in robotic surgeries [36]. Additionally, there are numerous other examples with varying sensor configurations documented in the literature, such as LMC-LMC [37], LMC-Microsoft Kinect [38]. Moreover, [39] has introduced a dedicated sensor fusion schema for accurate fingertip estimation using sensor-fitted gloves and LMC. Another Kalman filter data fusion strategy proposed by [40] in order to estimate the palm center by combining the position data gathered from LMC and velocity data obtained from Microsoft Kinect. Reference [40] suggested a Kalman filter data fusion approach to estimate the palm center using position data from LMC and velocity data from Microsoft Kinect.

B. Online Handover Location

A number of studies have been conducted to utilize the dynamic handover location oriented on the receiver's hand attitude. Dynamic Movement Primitives are used in the control schema of [41] to formulate a robot to follow human-like trajectories for object handover towards the human hand. Similarly, the object is delivered towards the human hand in [42], but this time using a proportional velocity controller. Another solution for dynamic handover location is introduced by [43], where an impedance-based control scheme is employed to drive the robot towards the predicted handover location. Reference [44] formulated the human-robot handover scenario by automatically synthesizing from high-level specifications in Signal Temporal Logic, and delivered the object by moving towards the human. In [45], the robot initially moves towards an expected handover location, and the trajectory is updated on the fly to converge smoothly to the actual handover location, which is estimated using optical motion capture that tracks optical markers on the delivery object.

C. Human Preferences

The consideration of human preferences during the handover process has been the subject of several studies. For

[1]Boston Dynamics https://www.bostondynamics.com/products/spot

[2]Microsoft Kinect https://developer.microsoft.com/en-us/windows/kinect

[3]Leap Motion https://www.leapmotion.com/

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

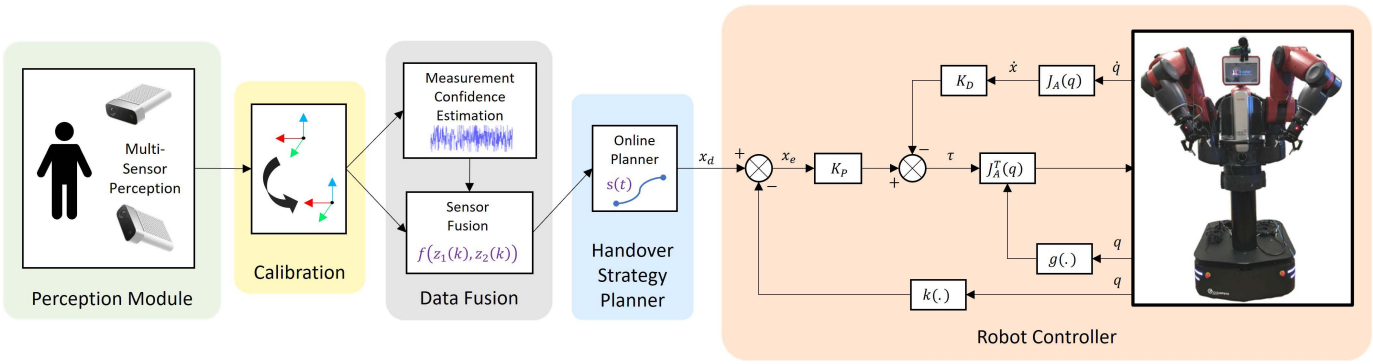


Fig. 3. Flow diagram of the proposed method including submodules of perception module, calibration, data fusion algorithms, online planner according to handover strategy, and robot control diagram.

example, a teaching-learning prediction framework is proposed in [1] for robots to learn from human partner demonstrations and predict human handover intentions by utilizing the human's preferences. Additionally, [46] presented a method that allows robots to continuously learn about the human preference regarding a specific task while working collaboratively with humans. An algorithm proposed in [47] aims to predict human intentions and use that information to plan robot trajectory and coordinate the human-robot collaboration. Another approach followed by [48] is to learn the optimal handover location for people with limited arm mobility. The researchers gathered data from 259 users to better understand their preferences in handover configuration by identifying each individual's most convenient handover configuration. In [2], a human-in-the-loop robotic system for collaborative assembly is studied by exploring the subtask distribution between robot and human.

D. Bimanual Object Handover

Although the literature has examined a range of realistic handover scenarios, bimanual object handovers have received less extensive investigation. A bidirectional bimanual human-robot handover system is designed in [3] for passing over a large planar object with a vertical posture. The researchers applied a position adjustment mechanism for the human-robot handover system using a dual-arm admittance neural network controller. For the human-to-robot object handover phase, they used QR codes on the object for localization. However, their perception system did not take advantage of the receiver's pose to hand over the object directly to the human operator's hands. Another comprehensive research in the domain of bimanual object handovers was conducted by [49], which adopted a learning-based approach to the problem. They utilized online updating probabilistic models to learn the beginning of the handover positions and tasks. They also benefited from two perception sensors fixed to the ground.

III. Methodology

In the following paragraphs, we present the perception module for extracting human pose, the calibration between sensor frames and robot frame, as well as the occlusion-resistant multi-sensor data fusion, including proposed noise estimation and data fusion techniques. We subsequently present the robot control strategy, and the handover strategy for involving human preferences in robot-to-human object handover. A flow diagram of this method with the mentioned submodules is depicted in Fig. 3.

A. Perception Module

In order to perform handover, the human pose is extracted via Microsoft Azure Kinect sensors. Microsoft Azure Kinect sensors are chosen for their remarkable adaptability to a broad range of environmental conditions, as well as their demonstrated efficacy in recognizing human postures, even when faced with partial occlusions [50]. Azure Kinect Software Development Kit (SDK) version 1.4 and Azure Kinect Body Tracking SDK version 1.1 with narrow field-of-view depth mode used as human pose extraction software. A skeleton overlaying the human pose is tracked in real-time containing 32 major joints.[4] The utilized body pose tracking algorithm works within operating range of 0.5 - 3.86 [m] with a random error standard deviation $\leq 17\,[mm]$, typical systematic error $< 11\,[mm] + 0.1\%$ of distance and it can run up to 30 FPS.[5]

The objective of this paper is to perform a bimanual handover directly to the hands of the human recipient. Therefore, we are interested in hand poses. The poses of hand measurements are notated with respect to the sensor's frame as ${}^{S,j}\boldsymbol{T}_{H,i}$ where $(H, i : i = \{left, right\})$ represents left and right hands, and $(S, j : j = \{1, 2\})$ denotes first and second sensors.

B. Calibration

The first step in the data fusion and sensor integration for HRI is the calibration of different reference frames. Each sensor produces measurements in its own reference frame, while the robot controller requires inputs in the robot's reference frame. Therefore, initially, the two Kinects are calibrated with respect to each other, and then hand-eye calibration is performed between the sensor and robot reference frames. These reference frames are represented in Fig. 4.

[4]https://docs.microsoft.com/en-us/azure/kinect-dk/body-joints
[5]https://docs.microsoft.com/en-us/azure/kinect-dk/hardware-specification

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

Fig. 4. Experimental setup for naturalistic object handover of the big-sized box bimanual handover from robot-to-human while the human is standing on the ladder, annotated with the defined coordinate systems.

1) Calibration Between Sensors: In order to calibrate the sensors, a two-minute data acquisition session is held while tracking the body pose from both sensors. During this session, the human operator freely moved in the working space and performed a variety of random motions. Position data of 32 joints are recorded from the two sensors, and pre-processing is done to combine the data when both sensors can detect body pose clearly. Then, this dataset is used to find the calibration matrix (${}^{S1}\boldsymbol{T}_{S2}$) between sensors by using Horn's Method [51]. This calibration matrix is used to calibrate the hand measurements from the first Kinect sensor (${}^{S1}\boldsymbol{T}_{H,i}$) and second Kinect sensor (${}^{S2}\boldsymbol{T}_{H,i}$). By using this calibration matrix, hand estimates from the second Kinect sensor are calibrated to the first Kinect sensor reference frame as follows:

$$ {}^{S1}\widehat{\boldsymbol{T}}_{H,i} = {}^{S1}\boldsymbol{T}_{S2}\ {}^{S2}\boldsymbol{T}_{H,i} \tag{1} $$

Finally, hand measurements from the first Kinect (${}^{S1}\boldsymbol{T}_{H,i}$) and the second Kinect calibrated to the first Kinect reference frame (${}^{S1}\widehat{\boldsymbol{T}}_{H,i}$) are used for data fusion as they refer to the same reference frame.

2) Hand-Eye Calibration: Although calibration between sensors is sufficient for data fusion, hand-eye calibration is needed to generate commands in the robot's reference frame. To achieve this, the depth vision (point cloud) generated by the first sensor is used. The point cloud is utilized to identify the corresponding point to the end effector position ($\boldsymbol{p}_{EE}^{S1}$), and the end effector position is also recorded using forward kinematics from the robot frame ($\boldsymbol{p}_{EE}^{R}$). 20 data points for hand-eye calibration are collected with this method, and Nelder-Mead method [52] is used to find the calibration matrix (${}^{R}\boldsymbol{T}_{S1}$) between robot and sensor 1 reference frame. Finally, hand estimations are transferred to robot frame (${}^{R}\boldsymbol{T}_{H,i}$) for the first and second Kinect, respectively, as:

$$ \begin{aligned} {}^{R}\widehat{\boldsymbol{T}}_{H,i} &= {}^{R}\boldsymbol{T}_{S1}\ {}^{S1}\boldsymbol{T}_{H,i} \\ {}^{R}\widehat{\boldsymbol{T}}_{H,i} &= {}^{R}\boldsymbol{T}_{S1}\ {}^{S1}\boldsymbol{T}_{S2}\ {}^{S2}\boldsymbol{T}_{H,i} \end{aligned} \tag{2} $$

C. Data Fusion

In this paper, a method to estimate noise in the body pose tracking data is proposed. This information is utilized to improve sensor fusion, especially in terms of resilience to occlusions. The noise estimation algorithm predicts the confidence of the measurements as a weight for the specific sensor data, where a higher weight indicates higher reliability.

1) Noise Estimation: To estimate the noise, two metrics are utilized: 1) tracking existence and 2) jerk of the data. The tracking existence part checks if the tracking data is available by controlling whether the received information is null or the same as the previous instance. This algorithm generates a parameter named ($occlusion^{S,j} \in \{0\},\{1\}$) where ($j = \{1,2\}$) is the sensor index. If tracking is available, then the corresponding tracking existence parameter will be zero ($occlusion^{S,j} = 0$) and vice versa, when body pose can not be generated, the corresponding parameter will be one ($occlusion^{S,j} = 1$).

The second part of the algorithm computes the jerk of the hand ($Jerk_{H,i}^{S,j}$) data for both sensors ($j = \{1,2\}$) and both hands ($i = \{1,2\}$). Then, the norm of the jerk data for each sensor ($\|Jerk_{H,i}^{S,j}\|$) is calculated and saturated as $\|\tilde{Jerk}_{H,i}^{S,j}\| \in [0,\alpha]$. α value is selected based on our experiences for the defined function $\|Jerk_{H,i}^{S,j}\|/\alpha \in [0,1]$ to be equal to 1 when there is a critical amount of high jerk occurs, indicating unreliable measurements. Jerk information in the data is used to give more weight to a sensor that generates more precise measurements, and even more crucially, flickering starts just before the occlusion happens, and it results in very high jerks. In this way, the estimation of an occlusion even before the SDK generates null information is possible, thus making the handover transfer more robust. Finally, noise estimation module generates weights ($w_j \in [0,1]$) for each sensor ($j = 1,2$) based on the ($occlusion^{S,j}$ and $Jerk_{H,i}^{S,j}$) metrics using the proposed function:

$$ w_j = |1 - max(\|\tilde{Jerk}_{H,i}^{S,j}\|/\alpha, occlusion^{S,j})| \tag{3} $$

2) Sensor Fusion: The hand pose data is extracted, and the measurement confidence is computed for each sensor in the previous noise estimation section. In this module, the utilization of weighted averaging is proposed to fuse the sensor data of the tracked joint positions with the computed sensor weights:

$$ \mathbf{p}_{H_i,fused}^{R} = \frac{w_1}{w_1+w_2}\mathbf{p}_{H_i,S1}^{R} + \frac{w_2}{w_1+w_2}\mathbf{p}_{H_i,S2}^{R} \tag{4} $$

This fusion equation gives more weight to the sensor with higher measurement confidence. But, more importantly, when there is an occlusion in one sensor, corresponding measurement confidence will return null ($w_j = 0$), and the sensor fusion algorithm will continue to estimate body pose using the data from the other sensor. Moreover, the solution will not be feasible if tracking data is not available ($w_1 + w_2 = 0$).

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

However, this would not affect our method because robot control will be enabled only if at least one sensor can track the human pose with a determined confidence threshold ($w_j > 0.5$).

Henceforth, the fused left hand palm position will be denoted as ($\boldsymbol{x}_{H,l}$), and the fused right hand palm position will be denoted as ($\boldsymbol{x}_{H,r}$) for simplicity in the next sections.

D. Robot Control

In this handover research, where bimanual object manipulation is performed and the handover location is updated during the handover phase, a reliable controller for the robot manipulators is required. Due to the nature of bimanual action, synchronized motion is essential for both arms with minimum error. For this reason, a PD + gravity Cartesian space controller and an improved orientation controller that benefits from the unit quaternion approach to endure singularities are utilized.

1) PD + Gravity Cartesian Space Controller: The robot dynamical model can be written as follows:

$$\boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} + \boldsymbol{g}(\boldsymbol{q}) - \boldsymbol{\tau}_e = \boldsymbol{\tau}_C \tag{5}$$

where $\boldsymbol{q} \in R^n$ is the joint values vector, $\boldsymbol{M}(\boldsymbol{q}) \in R^{n\times n}$ is the inertia matrix, $\boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}}) \in R^{n\times n}$ is a matrix representing the Coriolis and Centrifugal effects, and $\boldsymbol{g}(\boldsymbol{q}) \in R^n$ is the vector of gravity torques. $\boldsymbol{u} \in R^n$ represent the control torques The torque vectors $\boldsymbol{\tau}_C \in R^n$ and $\boldsymbol{\tau}_e \in R^n$ represent the control torques and the external disturbance torque vectors, respectively. Controller is designed based on the computed error ($\boldsymbol{e}$) for both position ($\boldsymbol{e}_p$) and orientation error ($\boldsymbol{e}_o$) as:

$$\boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}_p \\ \boldsymbol{e}_o \end{bmatrix} \tag{6}$$

where error ($\boldsymbol{e}_p$) is defined between desired position ($\boldsymbol{x}_d$) and current position ($\boldsymbol{x}_c$) of the end effector:

$$\boldsymbol{e}_p = \boldsymbol{x}_d - \boldsymbol{x}_c \tag{7}$$

Moreover, the orientation error is derived using unit quaternions, which are explained in detail in the next section. Finally, PD + gravity Cartesian space controller is designed as given in Fig. 3. The derivative term of the controller is only applied to output ($\dot{\boldsymbol{x}}_d = 0$) as in P-D structure to stabilize the motion. The corresponding control law can be written as follows:

$$\boldsymbol{\tau}_c = \boldsymbol{g}(\boldsymbol{q}) + \boldsymbol{J}^{+}(\boldsymbol{q})\boldsymbol{K}_p\, \boldsymbol{e} - \boldsymbol{J}^{+}(\boldsymbol{q})\boldsymbol{K}_d\dot{\boldsymbol{e}}_c \tag{8}$$

In the application case, the Baxter robot (Rethink Robotics, Bochum, Germany), which is robot with two 7-degrees-of-freedom (DOF) arms has been used. To handle kinematic redundancy, the right (Moore Penrose) pseudo-inverse of the Jacobian matrix ($\boldsymbol{J}^{+}(\boldsymbol{q})$) is used in the controller with the objective of optimizing the cost function of the joint velocities.

2) Orientation Controller - Unit Quaternion Approach: The orientation of the robot is defined with a rotation matrix by using 9 parameters and 6 orthonormal constraints, which satisfy orthogonality and unit length conditions [53]. However, there is a need to convert this rotation matrix to another controllable version. For this reason, first, we tested an Euler matrix representation, which could not achieve stable motion due to the singularities that are frequently present in our robot configuration. Therefore, an alternative definition is employed by resorting to a four-parameter singularity-free representation through unit quaternion. Unit quaternion ($U_q = \{\eta, \boldsymbol{\epsilon}\}$) is defined as:

$$\eta = \cos\theta/2 \tag{9}$$

$$\boldsymbol{\epsilon} = \sin\theta/2\, \boldsymbol{r} \tag{10}$$

where θ and $\boldsymbol{r}$ are respectively the rotation and the (3×1) unit vector of an equivalent angle/axis orientation description. In order to utilize quaternions to compute errors, a method introduced in [54] was applied. Orientation error ($\boldsymbol{e}_o \in R^3$) by utilizing unit quaternions is computed as:

$$\boldsymbol{e}_o = \eta_c(\boldsymbol{q})\boldsymbol{\epsilon}_d - \eta_d\boldsymbol{\epsilon}_c(\boldsymbol{q}) - \boldsymbol{\epsilon}_d \times \boldsymbol{\epsilon}_c(\boldsymbol{q}) \tag{11}$$

where η_c and $\boldsymbol{\epsilon}_c(\boldsymbol{q})$ denotes current orientation representation for unit quaternion ($U_{q,c} = \{\eta_c, \boldsymbol{\epsilon}_c\}$), and η_d and $\boldsymbol{\epsilon}_d$ denotes desired orientation representation for unit quaternion ($U_{q,d} = \{\eta_d, \boldsymbol{\epsilon}_d\}$), respectively. Once the orientation error is computed ($\boldsymbol{e}_o$), the orientation controller could be applied in accordance with the eq. (6) and (8).

E. Handover Strategy

In order to enhance the human experience, this paper proposes an adaptive handover strategy that takes into consideration the human preferences for handing over the box from robot-to-human. Additionally, the handover location is updated by employing the occlusion-robust perception module.

1) Human Preference Based Handover Strategy: Two modes are introduced to design a handover strategy considering the posture of the human partner:

A) When the hands are pointing upward, it indicates that the human operator wants to grasp the box from below. Therefore, the robot follows the handover strategy, where it delivers the box from above the hands.
B) When the hands are pointing downward, it indicates that the receiver wants to hold the box from above, and the robot executes the handover of the box from below the hands.

Handover location is determined based on the hands' pose information, the measurement confidence of the sensor, and the dimensions of the handover object. The mathematical notations are shown in Fig. 5.

To find the handover location, initially, the center ($\boldsymbol{x}_{H,c}$) of the left ($\boldsymbol{x}_{H,l}$) and right ($\boldsymbol{x}_{H,r}$) hands location is computed:

$$\boldsymbol{x}_{H,c} = \frac{\boldsymbol{x}_{H,r} + \boldsymbol{x}_{H,l}}{2} \tag{12}$$

Moreover, relative orientation of the hands is included via direction vector from left to right hand ($\boldsymbol{d}$), respectively:

$$\boldsymbol{d} = \frac{\boldsymbol{x}_{H,r} - \boldsymbol{x}_{H,l}}{\|\boldsymbol{x}_{H,r} - \boldsymbol{x}_{H,l}\|} \tag{13}$$

However, we have to constrain the z-axis as a simplification for keeping the balance of the box:

$$\boldsymbol{d}_{xy} = \begin{bmatrix} \boldsymbol{d}_x & \boldsymbol{d}_y & 0 \end{bmatrix} \tag{14}$$

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

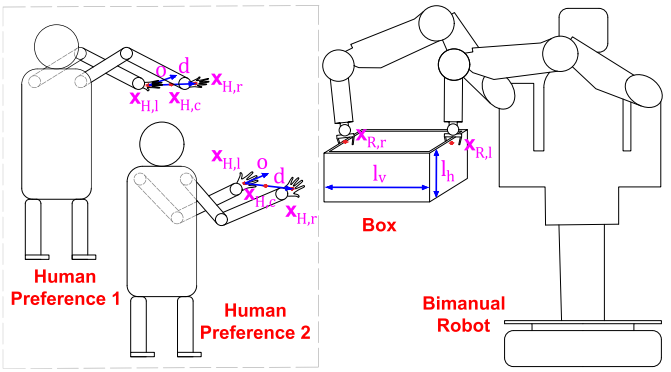


Fig. 5. Schema of the studied handover scenario with the annotated mathematical notations.

Correspondingly, the orthogonal unit vector between the z-axis of the ground ($\boldsymbol{k}$) and the constrained hand direction unit vector ($\boldsymbol{d}_{xy}$) is computed to find offset unit vector ($\hat{\boldsymbol{o}}$), and 10 [cm] offset is applied for the easiness of grasping:

$$\begin{aligned} \hat{\boldsymbol{o}} &= \boldsymbol{k} \times \boldsymbol{d}_{xy} \\ \boldsymbol{o} &= 10\,\hat{\boldsymbol{o}} \end{aligned} \tag{15}$$

To determine the handover location, horizontal length (l_h) and vertical length (l_v) of the box are considered. When hands are facing downwards, the handover location is calculated as:

$$\begin{aligned} \boldsymbol{x}_{R,l} &= \boldsymbol{x}_{H,c} + \boldsymbol{d}_{xy}\, l_h/2 + \boldsymbol{o} \\ \boldsymbol{x}_{R,r} &= \boldsymbol{x}_{H,c} - \boldsymbol{d}_{xy}\, l_h/2 + \boldsymbol{o} \end{aligned} \tag{16}$$

On the other hand, when hands are pointing upside, the handover location is computed as:

$$\begin{aligned} \boldsymbol{x}_{R,l} &= \boldsymbol{x}_{H,c} + \boldsymbol{d}_{xy}\, l_h/2 + \boldsymbol{o} + l_v \\ \boldsymbol{x}_{R,r} &= \boldsymbol{x}_{H,c} - \boldsymbol{d}_{xy}\, l_h/2 + \boldsymbol{o} + l_v \end{aligned} \tag{17}$$

Additionally, the handover orientation is adjusted accordingly. Before concluding the desired position ($\boldsymbol{x}_d$) for left and right robotic manipulators, motion planning is required to synchronize the motion of both arms to carry the object without causing damage. For this purpose, we follow the leader-follower strategy as employed in [55], where one manipulator is considered as the leader, and the other follows the trajectory acquired from the leader's trajectory. In order to derive the follower's trajectory, the transformation between the left and right robotic arm when the object is picked up (${}^{R_p,l}\boldsymbol{T}_{R_p,r}$) in the first phase is stored in the robot frame, where p index represents pickup, R is robot frame, l and r are left and right robotic arms. Using this information, the trajectory of the follower arm can be computed based on which manipulator is to be estimated.

A hand with higher confidence is used to determine which arm will be the leader. To illustrate this, if the right hand had higher measurement confidence, then the corresponding left robotic manipulator is determined to be a leader. Conversely, if the left hand had higher confidence, then the right robotic manipulator would be the leader. It is worth noting that hands' orientation information received by the perception module is utilized to determine the mode of the handover strategy, whereas the position information of the hands is utilized to define the handover location.

Algorithm 1 Pseudo Code of the Proposed Framework

function SENSOR FUSION CALLBACK(${}^{S,j}\boldsymbol{T}_{H,i}$, ${}^{S,j}\boldsymbol{T}_{H,i}$)
  ▷ *Read sensor data for hand poses* ${}^{S,j}\boldsymbol{T}_{H,i}$, ${}^{S,j}\boldsymbol{T}_{H,i}$ ◁
  $w_j \leftarrow f({}^{S,j}\boldsymbol{T}_{H,i})$ ▷ *Compute measurement confidences*
  if $min(w_j > 0.5)$ **then**
    $\mathbf{p}^R_{H_{i,fused}} \leftarrow f(w_j)$ ▷ *Fuse data using confidences*
  return $\boldsymbol{x}_{H,i} = \mathbf{p}^R_{H_{i,fused}}$
function PREPARE_ROBOT(handover_state)
  procedure ENABLE_ROBOT ()
  if handover_state == "initialize" **then**
    Pickup_object()
    ${}^{R_p,l}\boldsymbol{T}_{R_p,r} \leftarrow ({}^{R}\boldsymbol{T}_{R_p,l})^{-1}\,{}^{R}\boldsymbol{T}_{R_p,r}$ ▷ *Record transformation between arms for leader-follower method*
    Bring_object() ▷ *approach to human by mobile base*
while handover_in_action AND $min(w_j > 0.5)$ **do**
  if hands_looking_upward **then**
    $\boldsymbol{x}_{R,l}, \boldsymbol{x}_{R,r} \leftarrow f(\boldsymbol{x}_{H,c}, \boldsymbol{d}_{xy}, l_h/2, \boldsymbol{o}, {}^{R_p,l}\boldsymbol{T}_{R_p,r})$
  else if hands_looking_downward **then**
    $\boldsymbol{x}_{R,l}, \boldsymbol{x}_{R,r} \leftarrow f(\boldsymbol{x}_{H,c}, \boldsymbol{d}_{xy}, l_h/2, \boldsymbol{o}, l_v, {}^{R_p,l}\boldsymbol{T}_{R_p,r})$
  $\boldsymbol{s}(t) \leftarrow f(\boldsymbol{x}_{R,l}, \boldsymbol{x}_{R,r}, e)$ ▷ *Update online trajectory*
  if $min(w_j < 0.5)$ **then**
    $\boldsymbol{x}_{R,l}, \boldsymbol{x}_{R,r} \leftarrow return_base_position()$
    break ▷ *wait until hands are detected again*
  if $max(\|\boldsymbol{x}_{H,r} - \boldsymbol{x}_{R,l}\|, \|\boldsymbol{x}_{H,r} - \boldsymbol{x}_{R,l}\|) < 2[cm]$ **then**
    deliver_object()

2) Online Handover Location Update: One of the significant contributions of our approach is the online update of handover location throughout the process, even in the presence of occlusions. These occlusions are caused by the big handover item, manipulators, and complex environment. However, the proposed occlusion resilient multi-sensor data fusion module overcomes occlusions by utilizing improved multi-sensor fields of view. As a result, the robot can still follow and successfully deliver the item, even if the human partner changes location and orientation of the hands. Continuous estimations of hands pose information ($\boldsymbol{x}_{H,c}$) and measurement confidence weights (w_j) are acquired online to update the handover location. Algorithm 1 demonstrates the framework.

IV. Experiments

A. Experimental Setup

To simulate a naturalistic environment for object handover, a bimanual handover of a large object from robot-to-human is performed while the human partner stands on a ladder to place the box on a shelf and awaits the box from the robot. Since bimanual manipulation is required for carrying the object, it is not safe for humans to climb without holding the ladder supports. Also, the robot cannot directly place the object on the shelf due to working space limitations. Therefore, the robot should bring the box to the human to place it on the shelf. A box with dimensions of (45 × 35 x 25) [cm] is used as an object. Baxter robot (Rethink Robotics, Bochum, Germany),

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

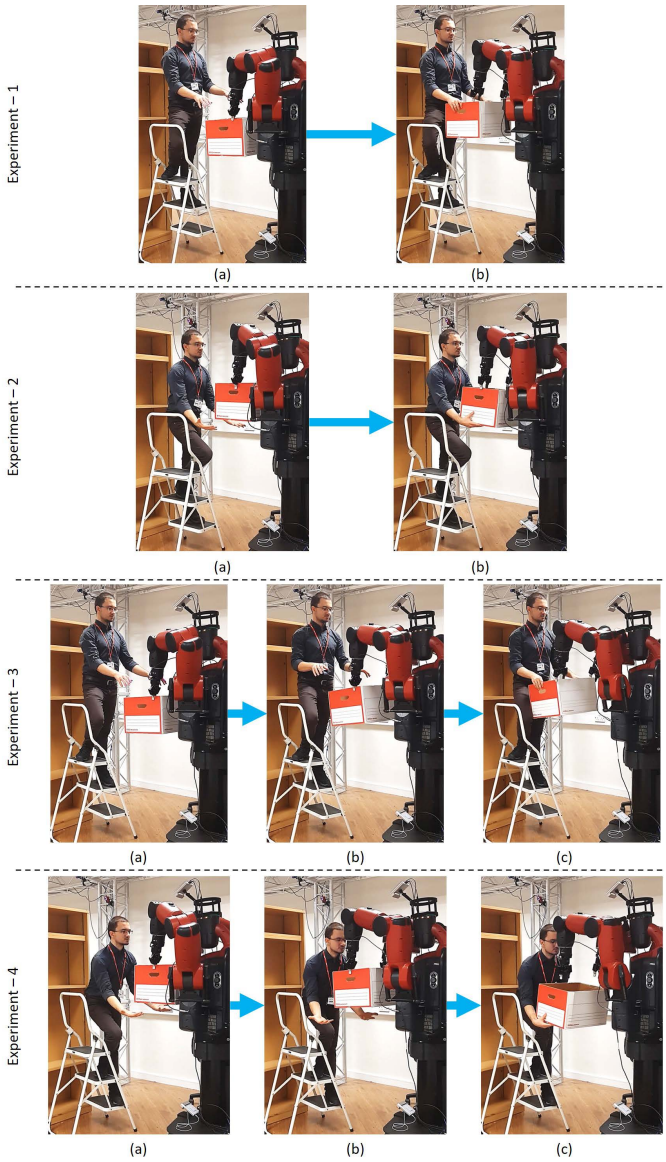


Fig. 6. Four sets of experiments while hands are facing Exp-1) downward, so the object is delivered beneath the hands; Exp-2) upward, so the object is delivered above the hands; Exp-3) similar approach with Exp-1 but handover location changed significantly during the operation; and Exp-4) similar approach with Exp-2 but handover location changed substantially during the procedure.

a dual-arm robotic manipulator, is used with a Ridgeback mobile base (Clearpath Robotics, Ontario, Canada) and two dual-finger grippers (Robotiq, Lévis, Canada).

B. Experiment Protocol

A set of four experiments involving different human preferences and significant handover location changes during the action were tested, as illustrated in Fig. 6.

1) *Downward*: receiver's hands are pointing downward, and correspondingly object is delivered from below the hands.
2) *Upward*: receiver's hands are pointing upward, and correspondingly object is delivered from above the hands.
3) *Location updated downward*: it starts the same with the *downward* case, but this time, the receiver changes the handover location in the middle of the process.

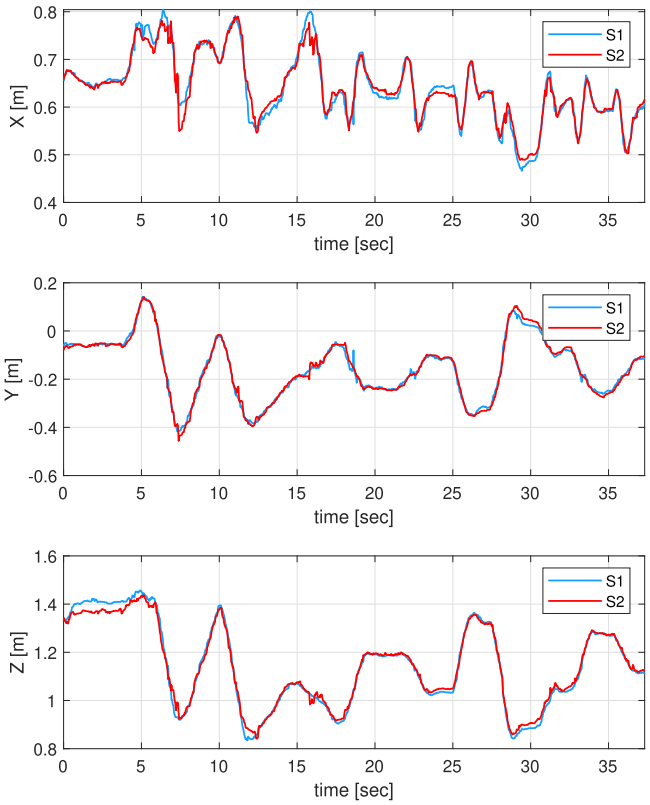


Fig. 7. Validation session trajectories for left hand's palm position estimated by sensor 1 (S1) and sensor 2 (S2) in robot's reference frame after the calibration between sensors (with an RMSE of 0.0241 [m]) and hand-eye calibration (with an RMSE of 0.0167 [m]). The likelihood of trajectories (RMS of Euclidean distance between calibrated measurements equals 0.0282 [m]) indicates successful calibration.

4) *Location updated upward*: it starts the same with the *upward* case, but this time, the receiver changes the handover location in the middle of the process.

Each task was performed for 30 repetitions, and a multi-sensor fusion technique was compared to a single sensor application of the top sensor (S1) and bottom sensor (S2), as well as a simple average method of both sensors. In total, 480 experiments were conducted, including these four scenarios.

V. Results

A. Calibration

Root mean square error (RMSE) of calibration between one sensor data and other calibrated sensor data is achieved as (0.0105,0.0095,0.0152) [m] in (x,y,z) and 0.0241 [m] in Euclidean distance. Calibrated sensor 1 data ($^{S1}\widehat{\boldsymbol{T}}_{H,i}$) and sensor 2 data ($^{S2}\boldsymbol{T}_{H,i}$) are presented in the Fig. 7 for (x,y,z).

On the other hand, RMSE of the hand-eye calibration between calibrated sensor estimation of end-effector position and measured end-effector position is acquired as (0.0050,0.0060, 0.0129) [m] in (x,y,z) and 0.0167 [m] in Euclidean distance.

B. Data Fusion

The data fusion algorithm estimates confidence values for sensors and combines this data to overcome occlusion problems mainly. To illustrate this, data related to the left hand

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

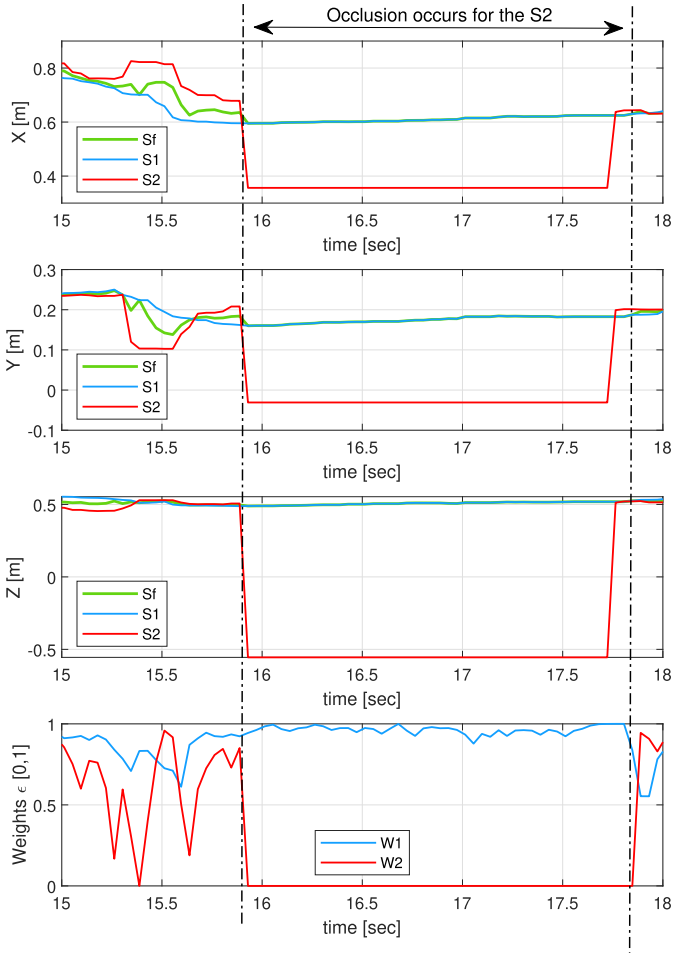


Fig. 8. The resulting estimations of the left hand's palm position during experiments from sensor 1 (S1), sensor 2 (S2), and sensor fusion estimates (Sf), followed by computed sensor weights (W1 and W2 for S1 and S2, respectively) for data fusion. In experiments, occlusion occurs from the S2 perspective, as indicated in the figure. Accordingly, the weight of the second sensor (W2) begins to signal unreliable estimations due to the presence of a high jerk even before occlusion occurs, and eventually, W2 reaches zero due to the occlusion. As a result, weighted average-based data fusion formulations rely more on the sensors with higher measurement confidence and provide continuous hand estimations even when one sensor is completely occluded.

position during occlusion occurs for sensor 2 (S2) are shown in Fig. 8. The figure plots the calibrated sensor data and the fused data regarding the left hand's palm position, along with computed weights for each sensor.

C. Robot Control

Initially, compliant motion with the leader-follower control structure is tested while carrying the box. A circular trajectory in space is given as input to the leader (left manipulator selected as a leader in this experiment), and the corresponding trajectory for the follower (right manipulator selected as a follower for this demonstration) is computed by using the fixed transformation matrix between two end effectors. The resulting trajectories are shown in Fig. 9. RMSE of the tracking performance is computed as 1.77 [cm] for the leader and 1.82 [cm] for the follower manipulator.

Furthermore, to illustrate the benefit of the handover location update, the results from experiment-3 are plotted in

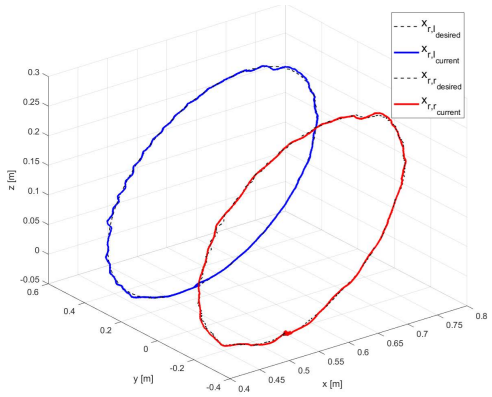


Fig. 9. Trajectory control results from the bimanual manipulation under compliant motion restrictions with a leader-follower control structure. The left manipulator (blue line) acts as the leader in this scenario, and the right manipulator (red line) is selected as the follower manipulator for this demonstration.

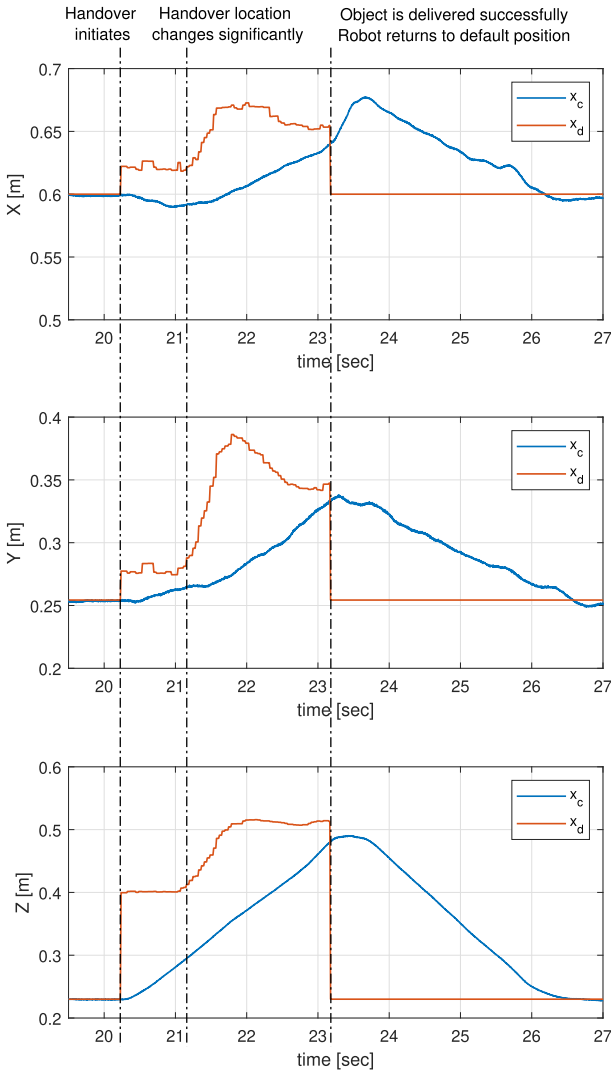


Fig. 10. Robot control trajectories for experiment-3 together with the desired (x_d) and current (x_c) position for the left robotic arm. After the handover was initiated, the handover location changed significantly during the action, requiring the robot to alter its trajectories for the handover to be successful.

Fig. 10 along with the desired and current position for the end effector of the left robotic arm. In this figure, it can be

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

TABLE I

HANDOVER PERFORMANCE COMPARISON OF THE FOUR SETS OF EXPERIMENTS FOR I) S1: SINGLE SENSOR INPUT OF THE FIRST SENSOR (TOP SENSOR), II) S2: SINGLE SENSOR INPUT OF THE SECOND SENSOR (BOTTOM SENSOR), III) AVG (S1, S2): SIMPLE AVERAGING OF DATA COLLECTED FROM BOTH SENSORS, IV) SF: PROPOSED MULTI-SENSOR FUSION SYSTEM. A HANDOVER SUCCESS RATE AND CORRESPONDING HANDOVER DURATION IN SECONDS WITH A 95% CONFIDENCE INTERVAL ARE REPORTED FOR 30 REPETITIONS PER SCENARIO (480 EXPERIMENTS IN TOTAL).BOLD HIGHLIGHTED BLOCKS REPRESENT THE STRENGTHS OF THAT TECHNIQUE

Method	**Handover Success Rate [%]**				**Handover Duration [sec]**			
	Exp - 1	Exp - 2	Exp - 3	Exp - 4	Exp - 1	Exp - 2	Exp - 3	Exp - 4
S1	**96.7**	16.7	**90.0**	0.0	**3.02 ± 0.33**	3.77 ± 2.50	**3.69 ± 0.28**	No Data
S2	26.7	**93.3**	0.0	**90.0**	7.20± 0.46	**4.09±1.39**	No Data	**3.84±0.90**
Avg (S1, S2)	**60.0**	**40.0**	6.7	0.0	**4.32± 0.65**	**5.39±0.8**	9.55±2.3	No Data
Sf	**96.7**	**90.0**	**96.7**	**86.7**	**3.18± 0.43**	**3.53 ± 0.99**	**3.88± 0.42**	**3.05 ± 0.63**

seen that the handover location moves from the initial location by approximately 0.1 [m] in the X direction, slightly in the Y direction, and 0.05 [m] in the Z direction of the robot reference frame. Furthermore, as shown in the accompanying figure, the object is delivered when the difference between the targeted handover location and the current robot position reaches a certain threshold.

D. Handover Experiments

For comparison of the proposed methodology, a bimanual object handover framework is tested with 1) only using a top sensor (S1), 2) only using a bottom sensor (S1), 3) simple averaging of both sensors, and 4) proposed sensor fusion. Handover success rate and handover duration are reported in TABLE I for each experiment and each method (30 repetitions for each case, 480 repetitions in total). It can be seen that the proposed sensor fusion algorithm brings the strengths of sensors from different perspectives and allows for the successful delivery of large-sized objects under occlusion.

VI. CONCLUSION AND FUTURE WORKS

In this article, we introduced a multi-sensor fusion technique supported by the proposed data fusion algorithms to achieve robot-to-human object handover in complex scenarios. Particularly, the case of handling a large object that requires bimanual manipulation in a naturalistic setup, such as when the operator is standing on a ladder, which is challenging in terms of perception limitations, is studied. To achieve this, a multi-sensor system using two depth cameras from different fields of view has been designed. Measurement confidence based on the jerk and data existence is computed to handle the occlusion in challenging environments. A data fusion algorithm is suggested to combine the sensor data with maximizing robustness. Finally, a Cartesian space controller is developed and utilized to complete the handover aim.

30 repetitions were performed for each of the 4 experiments (considering handover preferences and dynamic handover location) and 4 comparison methods, including raw single sensor input from the top sensor, raw single sensor input from the bottom sensor, simple averaging of the two sensor data, and the proposed multi-sensor fusion method. A simple averaging technique was able to combine the strengths of both sensors by sacrificing the handover success rate. However, it was not able to tolerate handover location alterations during the operation (Exp - 3 and Exp - 4), and also success rate was significantly lower than the proposed sensor fusion technique. Results showed that the proposed method fuses the strengths of both sensor fields of view, and accomplishes handover accuracy above 86.7% for all scenarios without sacrificing handover duration and adaptability to handover location adaptability.

This paper contributes to the literature by proposing a new sensor fusion framework that is resilient to occlusions, designing an online handover location incorporating the hands' location. The generic algorithms and frameworks introduced in this paper could be easily extended to different sensors, robots, and scenarios.

Limitations of this study are: i) The system depends on Azure Kinect's human pose detection capability since human pose tracking data is used as an input to the fusion system. Better utilization of different fields of view could be done by training a new human pose estimation algorithm based on the combined figure by utilizing advanced deep learning algorithms [56]. Another improvement could be made by uniting point clouds and using a point cloud-based detection algorithm [57]. ii) Success of the transfer phase is not detected. Object delivery is attempted when the object is 2 [cm] away from the desired point. Further work can be introduced to achieve smooth and guaranteed transfer, for example, utilizing force sensors. iii) Grasping technique is demonstrated only on this large object, and different objects will be tested in the subsequent work. Grasping techniques from recent computer vision studies can drastically improve the system's adaptability. In future works, we aim to study Bayesian methods to improve the precision of the data estimations. Additionally, various natural setups and other scenarios could be tested. Finally, more advanced control techniques, such as admittance control, could be utilized to improve the object handover experience.

ACKNOWLEDGMENT

The authors would like to thank Stelios Kotsovolis and Dr. Fan Zhang for their help and support in the implementation.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

REFERENCES

[1] W. Wang, R. Li, Y. Chen, Y. Sun, and Y. Jia, "Predicting human intentions in human–robot hand-over tasks through multimodal learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 3, pp. 2339–2353, Jul. 2022.

[2] M. Raessa, J. C. Y. Chen, W. Wan, and K. Harada, "Human-in-the-loop robotic manipulation planning for collaborative assembly," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 1800–1813, Oct. 2020.

[3] W. He, J. Li, Z. Yan, and F. Chen, "Bidirectional human–robot bimanual handover of big planar object with vertical posture," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 2, pp. 1180–1191, Apr. 2022.

[4] S. Kajikawa, T. Okino, K. Ohba, and H. Inooka, "Motion planning for hand-over between human and robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., Human Robot Interact. Cooperat. Robots*, Aug. 1995, pp. 193–199.

[5] R. Bischoff and T. Jain, "Natural communication and interaction with humanoid robots," in *Proc. 2nd Int. Symp. Humanoid Robots*, 1999, pp. 121–128.

[6] H. Su, A. Mariani, S. E. Ovur, A. Menciassi, G. Ferrigno, and E. De Momi, "Toward teaching by demonstration for robot-assisted minimally invasive surgery," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 2, pp. 484–494, Apr. 2021.

[7] S. Bibi, N. Anjum, and M. Sher, "Automated multi-feature human interaction recognition in complex environment," *Comput. Ind.*, vol. 99, pp. 282–293, Aug. 2018.

[8] V. Ortenzi, A. Cosgun, T. Pardi, W. P. Chan, E. Croft, and D. Kulic, "Object handovers: A review for robotics," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1855–1873, Dec. 2021.

[9] G. Pezzulo, F. Donnarumma, and H. Dindo, "Human sensorimotor communication: A theory of signaling in online social interactions," *PLoS ONE*, vol. 8, no. 11, Nov. 2013, Art. no. e79876.

[10] A. Mousavian, C. Eppner, and D. Fox, "6-DOF GraspNet: Variational grasp generation for object manipulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2901–2910.

[11] F. Cini, V. Ortenzi, P. Corke, and M. Controzzi, "On the choice of grasp type and location when handing over an object," *Sci. Robot.*, vol. 4, no. 27, Feb. 2019, Art. no. eaau9757.

[12] W. Yang, C. Paxton, A. Mousavian, Y. Chao, M. Cakmak, and D. Fox, "Reactive human-to-robot handovers of arbitrary objects," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 3118–3124.

[13] P. Rosenberger et al., "Object-independent human-to-robot handovers using real time robotic vision," *IEEE Robot. Autom. Lett.*, vol. 6, no. 1, pp. 17–23, Jan. 2021.

[14] J. Aleotti, V. Micelli, and S. Caselli, "An affordance sensitive system for robot to human object handover," *Int. J. Social Robot.*, vol. 6, no. 4, pp. 653–666, Nov. 2014.

[15] C.-M. Huang, M. Cakmak, and B. Mutlu, "Adaptive coordination strategies for human–robot handovers," in *Proc. Robot., Sci. Syst.*, vol. 11. Rome, Italy, 2015, pp. 1–10.

[16] A. C. Huamán Quispe, E. Martinson, and K. Oguchi, "Learning user preferences for robot-human handovers," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 834–839.

[17] A. Koene et al., "Relative importance of spatial and temporal precision for user satisfaction in human–robot object handover interactions," in *Proc. 3rd Int. Symp. New Frontiers Hum.-Robot Interact.*, 2014.

[18] Z. Han and H. Yanco, "The effects of proactive release behaviors during human–robot handovers," in *Proc. 14th ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, Mar. 2019, pp. 440–448.

[19] H. B. Suay and E. A. Sisbot, "A position generation algorithm utilizing a biomechanical model for robot–human object handover," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 3776–3781.

[20] H. Razali and Y. Demiris, "Multitask variational autoencoding of human-to-human object handover," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 7315–7320.

[21] S. Parastegari, E. Noohi, B. Abbasi, and M. Žefran, "A fail-safe object handover controller," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 2003–2008.

[22] A. G. Eguíluz, I. Rañó, S. A. Coleman, and T. M. McGinnity, "Reliable object handover through tactile force sensing and effort control in the shadow robot hand," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 372–377.

[23] N. Vahrenkamp et al., "Workspace analysis for planning human–robot interaction tasks," in *Proc. IEEE-RAS 16th Int. Conf. Humanoid Robots*, Nov. 2016, pp. 1298–1303.

[24] A. Kupcsik, D. Hsu, and W. S. Lee, "Learning dynamic robot-to-human object handover from human feedback," in *Robotics Research*. Berlin, Germany: Springer, 2018, pp. 161–176.

[25] E. C. P. Silva, E. W. G. Clua, and A. A. Montenegro, "Sensor data fusion for full arm tracking using Myo armband and leap motion," in *Proc. 14th Brazilian Symp. Comput. Games Digit. Entertainment (SBGames)*, Nov. 2015, pp. 128–134.

[26] C. H. Kang, C. G. Park, and J. W. Song, "An adaptive complementary Kalman filter using fuzzy logic for a hybrid head tracker system," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 9, pp. 2163–2173, Sep. 2016.

[27] X. Zhang and Y. Demiris, "Visible and infrared image fusion using deep learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 30, 2023, doi: 10.1109/TPAMI.2023.3261282.

[28] Y. Xu, X. Liu, Y. Liu, and S. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4256–4265.

[29] S. M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2006, pp. 133–146.

[30] B. Ives, K. Cossick, and D. Adams, "Amazon go: Disrupting retail?" *J. Inf. Technol. Teaching Cases*, vol. 9, no. 1, pp. 2–12, May 2019.

[31] S. Moon, Y. Park, D. W. Ko, and I. H. Suh, "Multiple Kinect sensor fusion for human skeleton tracking using Kalman filtering," *Int. J. Adv. Robotic Syst.*, vol. 13, no. 2, p. 65, Mar. 2016.

[32] A. Kitsikidis, K. Dimitropoulos, S. Douka, and N. Grammalidis, "Dance analysis using multiple Kinect sensors," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, vol. 2, Jan. 2014, pp. 789–795.

[33] S. Asteriadis, A. Chatzitofis, D. Zarpalas, D. S. Alexiadis, and P. Daras, "Estimating human motion from multiple Kinect sensors," in *Proc. 6th Int. Conf. Comput. Vis./Comput. Graph. Collaboration Techn. Appl.*, Jun. 2013, pp. 1–6.

[34] H. He, G. Liu, X. Zhu, L. He, and G. Tian, "Interacting multiple model-based human pose estimation using a distributed 3D camera network," *IEEE Sensors J.*, vol. 19, no. 22, pp. 10584–10590, Nov. 2019.

[35] S. E. Ovur, H. Su, W. Qi, E. De Momi, and G. Ferrigno, "Novel adaptive sensor fusion methodology for hand pose estimation with multileap motion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–8, 2021.

[36] W. Qi, S. E. Ovur, Z. Li, A. Marzullo, and R. Song, "Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 6039–6045, Jul. 2021.

[37] M. Mohandes, S. Aliyu, and M. Deriche, "Prototype Arabic sign language recognition using multi-sensor data fusion of two leap motion controllers," in *Proc. IEEE 12th Int. Multi-Conf. Syst., Signals Devices (SSD)*, Mar. 2015, pp. 1–6.

[38] B. Penelle and O. Debeir, "Multi-sensor data fusion for hand tracking using Kinect and leap motion," in *Proc. Virtual Reality Int. Conf.*, Apr. 2014, pp. 1–7.

[39] G. Ponraj and H. Ren, "Sensor fusion of leap motion controller and flex sensors using Kalman filter for human finger tracking," *IEEE Sensors J.*, vol. 18, no. 5, pp. 2042–2049, Mar. 2018.

[40] C. Li, A. Fahmy, and J. Sienz, "An augmented reality based human–robot interaction interface using Kalman filter sensor fusion," *Sensors*, vol. 19, no. 20, p. 4586, Oct. 2019.

[41] M. Prada, A. Remazeilles, A. Koene, and S. Endo, "Implementation and experimental validation of dynamic movement primitives for object handover," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 2146–2153.

[42] V. Micelli, K. Strabala, and S. Srinivasa, "Perception and control challenges for effective human-robot handoffs," in *Proc. RSS RGB-D Workshop*, Jun. 2011.

[43] J. R. Medina, F. Duvallet, M. Karnam, and A. Billard, "A human-inspired controller for fluid human–robot handovers," in *Proc. IEEE-RAS 16th Int. Conf. Humanoid Robots*, Nov. 2016, pp. 324–331.

[44] A. Kshirsagar, H. Kress-Gazit, and G. Hoffman, "Specifying and synthesizing human–robot handovers," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 5930–5936.

[45] M. K. X. J. Pan, E. Knoop, M. Bächer, and G. Niemeyer, "Fast handovers with a robot character: Small sensorimotor delays improve perceived qualities," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 6735–6741.

[46] T. Munzer, M. Toussaint, and M. Lopes, "Preference learning on the execution of collaborative human–robot tasks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 879–885.

IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.

[47] G. J. Maeda, G. Neumann, M. Ewerton, R. Lioutikov, O. Kroemer, and J. Peters, "Probabilistic movement primitives for coordination of multiple human–robot collaborative tasks," *Auto. Robots*, vol. 41, no. 3, pp. 593–612, Mar. 2017.

[48] P. Ardón et al., "Affordance-aware handovers with human arm mobility constraints," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3136–3143, Apr. 2021.

[49] Z. Yan, W. He, Y. Wang, L. Sun, and X. Yu, "Probabilistic motion prediction and skill learning for human-to-cobot dual-arm handover control," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 30, 2022, doi: 10.1109/TNNLS.2022.3182973.

[50] C. Neupane, A. Koirala, Z. Wang, and K. B. Walsh, "Evaluation of depth cameras for use in fruit localization and sizing: Finding a successor to Kinect v2," *Agronomy*, vol. 11, no. 9, p. 1780, Sep. 2021.

[51] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 4, no. 4, pp. 629–642, 1987.

[52] F. Gao and L. Han, "Implementing the Nelder–Mead simplex algorithm with adaptive parameters," *Comput. Optim. Appl.*, vol. 51, no. 1, pp. 259–277, Jan. 2012.

[53] F. Caccavale, C. Natale, B. Siciliano, and L. Villani, "Resolved-acceleration control of robot manipulators: A critical review with experiments," *Robotica*, vol. 16, no. 5, pp. 565–573, Sep. 1998.

[54] S. Chiaverini and B. Siciliano, "The unit quaternion: A useful tool for inverse kinematics of robot manipulators," *Syst. Anal. Model. Simul.*, vol. 35, no. 1, pp. 45–60, 1999.

[55] D. Rakita, B. Mutlu, M. Gleicher, and L. M. Hiatt, "Shared control-based bimanual robot manipulation," *Sci. Robot.*, vol. 4, no. 30, May 2019, Art. no. eaaw0955.

[56] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 1–41, Apr. 2023.

[57] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and robust hand tracking from depth," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1106–1113.

Salih Ertug Ovur (Graduate Student Member, IEEE) received the B.Sc. degree in mechanical engineering and the degree in control and automation engineering from Istanbul Technical University, Turkey, and the M.Sc. degree in automation and control engineering from Politecnico di Milano, Italy. He is currently pursuing the Ph.D. degree in improving safety and ergonomics in human–robot interaction through multimodal sensor fusion. During his B.Sc. degree, he was a Teaching and Research Assistant with the Artificial Intelligence and Intelligent Systems Laboratory, where he developed multiple theses on the designing, manufacturing, sensing, and control of mobile robots. He was a part-time Research and Teaching Assistant with the NEARLAB's Medical Robotics Section during his two-year M.Sc. studies and subsequently for a year as a full-time Research Collaborator. He used multi-sensor fusion to develop new frameworks for human–robot interaction and collaboration. His research is fully funded by Imperial College London President's Ph.D. Scholarships.

Yiannis Demiris (Senior Member, IEEE) received the B.Sc. degree (Hons.) in artificial intelligence and computer science and the Ph.D. degree in intelligent robotics from the Department of Artificial Intelligence, The University of Edinburgh, Edinburgh, U.K., in 1994 and 1999, respectively. He is currently a Professor with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., where he is also the Royal Academy of Engineering Chair of Emerging Technologies and the Head of the Personal Robotics Laboratory, and the Intelligent Systems and Networks Group. His current research interests include human–robot interaction, machine learning, user modeling, and assistive robotics. He has published more than 250 journals and peer-reviewed conference papers in the above areas. He was a recipient of the Rector's Award for Teaching Excellence in 2012 and the FoE Award for Excellence in Engineering Education in 2012. He is a fellow of the IET.