

# Benchmark for Human-to-Robot Handovers of Unseen Containers with Unknown Filling

Ricardo Sanchez-Matilla<sup>1</sup>, Konstantinos Chatzilygeroudis<sup>2</sup>, Apostolos Modas<sup>3</sup>,  
Nuno Ferreira Duarte<sup>2,4</sup>, Alessio Xompero<sup>1</sup>, Pascal Frossard<sup>3</sup>, Aude Billard<sup>2</sup>, and Andrea Cavallaro<sup>1</sup>

**Abstract**—The real-time estimation through vision of the physical properties of objects manipulated by humans is important to inform the control of robots for performing accurate and safe grasps of objects handed over by humans. However, estimating the 3D pose and dimensions of previously unseen objects using only RGB cameras is challenging due to illumination variations, reflective surfaces, transparencies, and occlusions caused both by the human and the robot. In this paper, we present a benchmark for dynamic human-to-robot handovers that do not rely on a motion capture system, markers, or prior knowledge of specific objects. To facilitate comparisons, the benchmark focuses on cups with different levels of transparencies and with an unknown amount of an unknown filling. The performance scores assess the overall system as well as its components in order to help isolate modules of the pipeline that need improvements. In addition to the task description and the performance scores, we also present and distribute as open source a baseline implementation for the overall pipeline to enable comparisons and facilitate progress.

**Index Terms**—Performance Evaluation and Benchmarking, Perception for Manipulation, Human-Robot Handover.

## I. INTRODUCTION

OVER the past decade, important progress has been made toward introducing robots in human-inhabited environments, such as factories and households. When robots and humans safely share a room or workspace, robots can be deployed to help people perform household chores, lift heavy loads or assist nurses in supporting the elderly. In this context, the capability of robots to exchange a wide range of objects with humans is particularly important.

Manuscript received: August, 15, 2019; revised November, 18, 2019; accepted December, 15, 2019. This paper was recommended for publication by Editor H. Ding upon evaluation of the Associate Editor and Reviewers' comments. This work is supported by the CHIST-ERA program through the project CORSMAL, under UK EPSRC grant EP/S031715/1 and Swiss NSF grant 20CH21\_180444; and the Research and Innovation program ICT-2014-1, under grant agreement 643950-SecondHands. N. Duarte is supported in part by a FCT-IST fellowship.

<sup>1</sup> R. Sanchez-Matilla, A. Xompero and A. Cavallaro are with Centre for Intelligent Sensing, Queen Mary University of London, UK {ricardo.sanchezmatilla, a.xompero, a.cavallaro}@qmul.ac.uk.

<sup>2</sup> K. Chatzilygeroudis, N. Ferreira Duarte and A. Billard are with LASA, Swiss Federal Institute of Technology, Lausanne, Switzerland {konstantinos.chatzilygeroudis, nuno.ferreiraduarte, aude.billard}@epfl.ch.

<sup>3</sup> A. Modas and P. Frossard are with LTS4, Swiss Federal Institute of Technology, Lausanne, Switzerland {apostolos.modas, pascal.frossard}@epfl.ch.

<sup>4</sup> N. Ferreira Duarte is with VisLab, Institute of Systems and Robotics, Lisbon, Portugal nferreiraduarte@isr.tecnico.ulisboa.pt.

R. Sanchez-Matilla and K. Chatzilygeroudis are co-first authors.

Digital Object Identifier (DOI): see top of this page.

Humans manipulate daily objects made of a wide variety of materials, thus making their safe and accurate handover to robots a challenging task, especially for previously unseen objects whose dimensions, stiffness, and weight are unknown. When humans are about to receive previously unseen objects from others, they first estimate the properties of the objects through vision and then use tactile and force feedback to improve this estimation. Even though the manipulation capabilities of robots exceed those of humans in terms of accuracy, their capabilities in terms of perception and dexterity still fall way behind those of humans. Open challenges include accurately estimating the object pose while the human moves the object, selecting the most appropriate grasping regions that will not harm the human, and predicting the filling of containers, such as boxes and cups, to estimate their mass and stiffness.

While vision has made considerable progress in the direction of supporting robotic manipulation tasks, most solutions exploit databases of 3D object models [1][2][3][4][5] or markers for motion capture [6]. To stimulate and support the development of algorithms for accurate and safe human-to-robot handovers without prior knowledge of the specific object characteristics or expensive setups for perception, we propose a benchmark for robots observing the manipulation of an object by a human to infer how to safely grasp the object when the human hands it over (Fig. 1). To facilitate comparisons, we consider as objects cups with different rigidities, dimensions, transparencies, shapes, and fillings. The proposed benchmark<sup>1</sup> aims to assess the generalization capabilities of the robotic control when handing over previously unseen objects filled with unknown content, hence with different and unknown mass and stiffness. No object properties are initially known to the robot, which must infer these properties on-the-fly, during the execution of the dynamic handover, through multi-modal perception of the scene. The evaluation is performed through an overall score to rank solutions as well as partial scores to determine differences between methods and provide guidance for improving individual elements of the system. To encourage and facilitate research in this area, we also provide a baseline solution and discuss its evaluation through the proposed benchmark in two different setups.

<sup>1</sup>The benchmark is an outcome of the project CORSMAL, which explores the fusion of sensing modalities (touch, sound, vision) to accurately and robustly estimate the physical properties of objects in noisy and potentially ambiguous environments: <http://corsmal.eecs.qmul.ac.uk/benchmark.html>. The webpage contains the links to the protocol and benchmark documents.

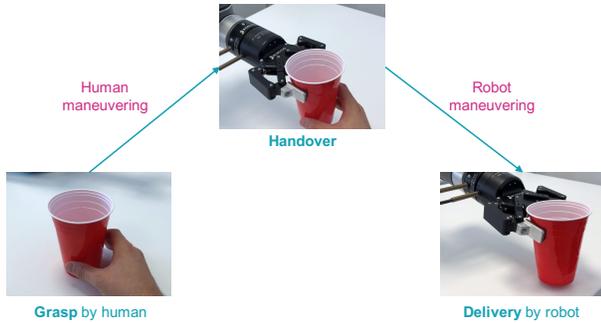


Fig. 1. The main phases of the human-to-robot handover of a cup.

## II. BACKGROUND

### A. Benchmarks

To address the need for replicating experiments and comparing solutions, the interest in benchmarks and shared datasets has recently increased considerably [7][8][9]. For example, the Yale - Carnegie Mellon University - Berkeley (YCB) object model and set [7] provides a database of objects with different shapes, sizes, textures, weights, and rigidities as well as a benchmark protocol with performance measures for tasks such as pouring a liquid from a container, pick-and-place for table setting, peg-insertion, and box-and-block tests (i.e. grasping and moving an object from a side of a box to another in a fixed amount of time). The Multimodal Grasp Dataset [9] uses visual and tactile data for a dexterous robotic hand to better understand the grasping process. RGB-D cameras generate point clouds of different objects placed at the center of a table in front of the robot. VisGraB [8] is a database for benchmarking approaches for vision-based grasping of unknown objects in realistic, everyday environments without the availability of prior object models. A software tool also allows researchers to replicate a combination of real-world and simulated experimental setups. For a comprehensive review of benchmarks and datasets for robot manipulation tasks, we refer the reader to [7].

### B. Handovers

Fluid and efficient human-to-robot handovers are the result of the combination of perception and control [10]. To this end, the pose of the object to hand over has to be tracked through vision to facilitate the planning and reaching of the predicted handover location, and the subsequent grasping and delivering of the object to the desired location.

Vision approaches to track the object pose and/or estimate its physical properties can be marker-based, model-based, or inference-based. Marker-based approaches [6][11] rely on motion capture systems to accurately track in 3D the pose of a human hand or object [12]. These approaches are expensive and require objects to be equipped with markers. Model-based approaches exploit prior knowledge on the object and its physical properties through dense 3D models [1][2][13][14]. Alternatively, datasets can be used to provide the annotation of grasp points on objects [15][16][17][18]. Large amounts

of annotated data facilitate the training of deep neural networks to estimate the object pose [1][2][13][14]. However, (manual) annotation is challenging as objects can be grasped in multiple ways, while annotators might be biased by the semantics of the object [19]. Moreover, this approach limits operations to *known* objects. With previously *unseen* objects, inference-based approaches localize them in 3D using only partial knowledge, such as dimensions, estimated through unimodal [20] or multi-modal sensing [21], or by estimating the 6 degrees of freedom (DoF) of textured objects by solving the Perspective-N-Point problem with known 2D-3D correspondences [22][23]. However, typical objects with textureless, reflective or transparent materials, like those in our benchmark, hinder reliable correspondences [24]. Moreover, depth sensors [25] cannot be applied with sufficient accuracy [15] or speed for a smooth handover [10].

Robots should also be able to quickly adapt their grip force during the handover. The design of solutions for human-to-robot handovers can be informed by the behavior of human-to-human interactions in handovers through models of the relationship between the amount of force applied on the object by each party (grip force) and the proportion of the object's weight carried by each party (load force) [10][26][27][28]. When modelling the motions between parties, solutions either find a complete trajectory (plan-based) [29][30][31] or continuously optimize the control output given the current state (controller-based) [32][33]. However, even if controllers can generate natural motions, robots may stop moving right before transitioning to the passing phase, once both parties make contact with the object. A combination of force modulation and online adaptation of hand closure and opening, inspired by human-to-human handovers, can achieve a fluid human-to-robot handover, but requires a marker-based tracking system and prior information about the object's content [6].

## III. THE BENCHMARK

### A. Setup and task

The setup for the benchmark consists of a robotic arm with a gripper, a table, cups and filling, and two cameras<sup>2</sup>. Participants are requested to submit a detailed description of the sensors and setup, including any controlled artificial illumination used during the experiments. Note that for this benchmark, a *participant* is any group or individual who evaluate their method(s) on the benchmark and submit some results, while a *subject* is any individual who is asked by the participants to perform the handover tasks.

At least four naive subjects (i.e. people external to the design and development of the system) perform the handover tasks for the benchmark. Each task consists of three phases, namely human maneuvering, dynamic handover, and robot maneuvering (Fig. 1). All the handover tasks of one subject shall be executed using the same hand. Subjects are instructed to handover the cup naturally, standing opposite to the robot across the table, which should be covered with a white tablecloth. The subject grasps the cup from a pre-defined location at the center of

<sup>2</sup>In addition, a third camera to record the overall scene shall be used for evaluation purposes only.

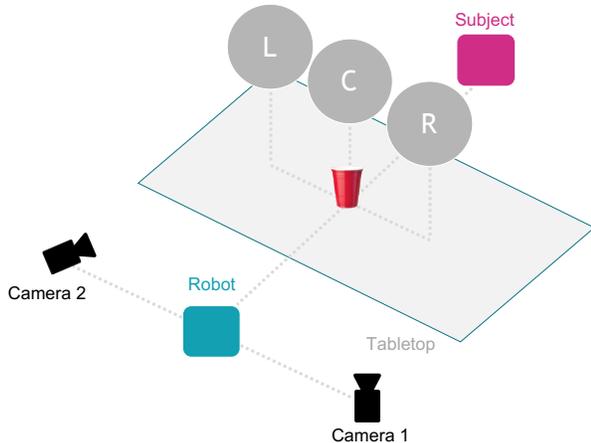


Fig. 2. The setup for the benchmark and the three handover locations (left, L; center, C; right, R), which are defined to be reachable by the robotic arm and comfortable for the subjects.

the table and carries the cup to one of the three approximate handover locations above the table: left, center, and right with respect to the robot (Fig. 2). As one cannot explicitly set these locations, assuming the subjects are executing the handover naturally (i.e. with no intention to help or make it difficult for the robot), these locations can be roughly constrained by the reachability of the arm (i.e. 40% – 50%), such that they are reachable by the robotic arm, but also comfortable for the subjects to perform the handover naturally. After the handover, the robot places the object on the table at a pre-defined location.

Participants can use any robotic arm with at least 6 degrees of freedom (e.g. KUKA, UR5) equipped with a gripper (e.g. a 2-finger gripper or a robotic hand). Touch and pressure sensors can be used on the gripper to facilitate the inference of the physical properties of the object. Fig. 2 illustrates the setup for the benchmark.

We select four objects that are challenging for both perception and robotics due to the high variability of their physical properties (e.g. material, shape, texture and mass). Specifically, we select four drinking cups (see Fig. 3) that are available worldwide<sup>3</sup>: Cup 1, with high deformability and medium transparency; Cup 2, with average deformability and low transparency; Cup 3, with average deformability and high transparency; and Cup 4, the plastic wine glass from the YCB object database [34], which is not deformable and has high transparency. Moreover, we vary the mass and stiffness of each cup by filling it with rice, which is easy to purchase and - unlike liquids - harmless for the hardware in case of spilling. Table I summarizes the properties of the cups and their filling.

Each subject shall perform the handover with three different grasps in random order: Grasp 1 from the bottom of the cup, Grasp 2 from the top, and Grasp 3 naturally. While Grasps 1



Fig. 3. The cup and human-grasp types for the benchmark.

TABLE I  
THE PROPERTIES OF THE FOUR CUP TYPES AND THE FILLING AMOUNT FOR THE BENCHMARK.

Properties	Unit	Cup 1	Cup 2	Cup 3	Cup 4
Deformability	-	High	Medium	Medium	None
Transparency	-	Medium	Low	High	High
Width at the top	cm	7.2	9.7	9.9	8.0
Width at the bottom	cm	4.3	6.1	6.5	6.5
Height	cm	8.2	12.1	13.6	13.5
Weight	g	2.0	10.0	9.0	134.0
Volume	ml	179.0	497.0	605.0	354.0
Filling amount	ml	125.0	400.0	450.0	300.0

and 2 make it easier for the robot to grasp the cup, in Grasp 3 the hand of the subject may cover most of the surface of the cup, considerably reducing the candidate grasping regions (see Fig. 3).

In summary, the combination of 4 cup types, 2 filling amounts, 4 subjects, 3 grasp types, and 3 handover locations leads to a total of  $N = 288$  configurations. The approximate time to execute all configurations is 4 hours.

Note that more than four subjects can take part in the experiments if participants wish to have a larger pool of configurations. If participants perform experiments with a setup that differs from the one in Fig. 2, the results will be reported in a different portion of the leaderboard. Moreover, while participants can use any training data to refine vision and robotics algorithms, no data from the specific objects and filling used for the benchmark can be used (i.e. we do not

<sup>3</sup>The links to purchase all objects are listed in <http://cor-smal.eecs.qmul.ac.uk/benchmark/resources/RAL-SI-2020-P19-0835-V1.0.pdf>.

TABLE II

PERFORMANCE SCORES OF THE BENCHMARK AND RESULTS FOR THE BASELINE IN SETUP 1 (S1) AND SETUP 2 (S2). FOR EACH MEASURE  $a$  THE CORRESPONDING GROUND TRUTH IS  $\hat{a}$ . ALL THE SCORES ARE NORMALIZED. SCORES FOR MEASURES THAT ARE NOT COMPUTED ARE DENOTED WITH .00\*. DARK GRAY SHADED ROWS HIGHLIGHT THE THREE MAIN SCORES OF THE OVERALL TASK AND THE TWO SUBSYSTEMS, NAMELY VISION AND ROBOT. LIGHT GRAY SHADED ROWS HIGHLIGHT TWO SCORES THAT ARE COMPUTED ONCE OFFLINE, PRIOR TO THE EXECUTION OF THE TASK.

Group	Description	Unit	Measure	Score	Baseline results	
					S1	S2
Vision	Width at top	mm	$w^i$	$s_1 = \frac{1}{N} \sum_{i=1}^N \sigma_1(w^i, \hat{w}^i)$	.59	.59
	Width at bottom	mm	$w_b^i$	$s_2 = \frac{1}{N} \sum_{i=1}^N \sigma_1(w_b^i, \hat{w}_b^i)$	.55	.61
	Height	mm	$h^i$	$s_3 = \frac{1}{N} \sum_{i=1}^N \sigma_1(h^i, \hat{h}^i)$	.54	.58
	Mass (cup + filling)	g	$m_v^i$	$s_4 = \frac{1}{N} \sum_{i=1}^N \sigma_1(m_v^i, \hat{m}^i)$	.00*	.00*
	Fullness	%	$f^i$	$s_5 = \frac{1}{N} \sum_{i=1}^N (1 -  f^i - \hat{f}^i /100)$	.00*	.00*
	Vision score			$S_v = \sum_{j=1}^5 \lambda_j s_j$	.19	.20
Robot	Mass (cup + filling)	g	$m_r^i$	$s_6 = \frac{1}{N} \sum_{i=1}^N \sigma_1(m_r^i, \hat{m}^i)$	.00*	.00*
	Human-hand pose prediction	(mm, °)	$\mathbf{P}_t^l$	$s_7 = \sigma_3(\mathcal{P}, \hat{\mathcal{P}}, \varepsilon)$	.00*	.00*
	End-effector	(mm, °)	$\mathbf{E}^l$	$s_8 = \sigma_3(\mathcal{E}, \hat{\mathcal{E}}, \varepsilon)$	.94	.94
	Robot score			$S_r = \sum_{j=6}^8 \lambda_j s_j$	.31	.31
Task	Delivery location	mm	$d^i$	$s_9 = \frac{1}{N} \sum_{i=1}^N \sigma_2(d^i, \delta)$	.47	.37
	Mass of the delivered filling	g	$\omega^i$	$s_{10} = \frac{1}{N} \sum_{i=1}^N \sigma_1(\omega^i, \hat{\omega}^i)$	.49	.58
	Human maneuvering time	ms	$t_{hm}^i$	$s_{11} = \frac{1}{N} \sum_{i=1}^N \sigma_2(t_{hm}^i, \tau)$	.41	.03
	Handover time	ms	$t_{ho}^i$	$s_{12} = \frac{1}{N} \sum_{i=1}^N \sigma_2(t_{ho}^i, \tau)$	.46	.44
	Robot maneuvering time	ms	$t_{rm}^i$	$s_{13} = \frac{1}{N} \sum_{i=1}^N \sigma_2(t_{rm}^i, \tau)$	.45	.09
	Task score			$S_g = \sum_{j=9}^{13} \lambda_j s_j$	.46	.40
Benchmark score				$S = \frac{1}{3}(S_v + S_r + S_g)$	.32	.33

allow learning across configurations).

### B. Performance scores

We quantify the performance of the vision subsystem, the robotic subsystem, and the completion of the overall task.

For Vision, we measure the accuracy of the estimation of the *dimensions*, *mass* and *fullness* of the cup. Specifically, for each configuration  $i$ , we consider the estimated height,  $h^i$ , of the cup, its width at the top,  $w^i$ , its width at the bottom,  $w_b^i$  (in millimeters); its mass,  $m_v^i$ , including its filling (in grams); and its *fullness*,  $f^i$ , as percentage of filled volume of the cup.

For Robot, we measure the accuracy of the controller or sensors in estimating the *mass* of the cup,  $m_r^i$ , including its content; the *human-hand pose prediction* error; and the accuracy of the *end-effector* reaching accuracy. Participants shall compute the last two scores offline and only once to evaluate the prediction algorithm and the robotic controller. To this end, we provide pre-recorded data<sup>4</sup> of six hand poses trajectories,  $\hat{\mathcal{P}} = \{\hat{\mathbf{P}}_t^l : t = 0, \dots, T - 1; l = 0, \dots, 5\}$ , where  $T$  is the duration of the trajectory, and six desired end-effector poses,  $\hat{\mathcal{E}} = \{\hat{\mathbf{E}}^l : l = 0, \dots, 5\}$ . The algorithm used by the participants should predict and report the pose of the human hand,  $\mathbf{P}_t^l$ , for each future time  $t$  based on the current end-effector position of the robot and the current human hand pose, for human-hand pose prediction. Participants should also report the actual end-effector's pose,  $\mathbf{E}^l$ , after the robotic controller tries to reach the given end-effector poses.

<sup>4</sup>Data for offline scores available at <https://github.com/CORSMAL/Benchmark/tree/master/offlineScores>.

For the overall Task, we measure the total *execution time*, the amount of *delivered* filling and the accuracy of the *delivery* of the cup. To facilitate the computation of the corresponding score, the location for the delivery of the cup is the same as its initial location. We consider the distance,  $d^i$ , between the position of the center of the base of the cup at the end of the task with respect to the initial position (in millimeters), the mass of filling,  $\omega^i$ , that was spilled during the execution (in grams), and the time elapsed to complete the task. This elapsed time is divided in three temporal segments, namely the time between when the human gets in contact with the cup and when the robot gets in contact with the cup (the *human maneuvering time*,  $t_{hm}^i$ ), the time when the human and robot are simultaneously in contact with the cup (the *handover time*,  $t_{ho}^i$ ), and the time between when the human is last in contact with the cup and when the robot is last in contact with the cup after delivery (the *robot maneuvering time*,  $t_{rm}^i$ ). If the setup is not designed to report these elapsed times automatically, participants can manually annotate the three segments of the execution time offline by watching a recorded video for each configuration.

When the ground truth,  $b$ , is available for measure  $a \in \{w^i, w_b^i, h^i, m_v^i, m_r^i, \omega^i\}$ , we use as normalization function  $\sigma_1(\cdot, \cdot)$ :

$$\sigma_1(a, b) = \begin{cases} 1 & \text{if } b = a = 0 \\ 1 - \frac{|a-b|}{b} & \text{if } |a-b| < b \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where the first condition refers to configurations with no filling

for  $m_v^i$ .

When there is no ground-truth information, we use for measure  $a \in \{d^i, t_{\text{hm}}^i, t_{\text{ho}}^i, t_{\text{rm}}^i\}$  as normalization function  $\sigma_2(\cdot, \cdot)$ :

$$\sigma_2(a, \eta) = \begin{cases} 1 - \frac{a}{\eta} & \text{if } a < \eta \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where  $\eta \in \{\delta, \tau\}$  is the threshold that defines when an algorithm is unsuccessful for that measure. Specifically, we use as values  $\delta = 500$  mm and  $\tau = 5000$  ms.

With reference to Table II, we compute the score,  $s_j \in [0, 1]$  (where 1 is best), across all the configurations  $i$  for each measure as follows:

$$s_j = \begin{cases} \frac{1}{N} \sum_{i=1}^N \sigma_1(a, b) & \text{if } a \in \{w^i, w_b^i, h^i, m_v^i, m_r^i, \omega^i\} \\ \frac{1}{N} \sum_{i=1}^N \sigma_2(a, \eta) & \text{if } a \in \{d^i, t_{\text{hm}}^i, t_{\text{ho}}^i, t_{\text{rm}}^i\} \\ \frac{1}{N} \sum_{i=1}^N \left(1 - \frac{|a-b|}{100}\right) & \text{if } a = f^i, \end{cases} \quad (3)$$

with  $j = \{1, \dots, 13\}$ ,  $j \neq 7, 8$ .

For the human-hand pose prediction, we compute the score between the estimated trajectories,  $\mathcal{P}$ , and the ground-truth trajectories,  $\hat{\mathcal{P}}$ , as

$$\sigma_3(\mathcal{P}, \hat{\mathcal{P}}, \varepsilon) = \frac{\sum_{l=0}^5 \sum_{t=0}^{T-1} (\sigma_2(\epsilon_{t,l}, \varepsilon) + \sigma_2(\alpha_{t,l}, \frac{\pi}{2}))}{12T}, \quad (4)$$

where  $\epsilon_{t,l} \in [0, \infty)$  is the Euclidean distance for translation, and  $\alpha_{t,l} \in [0, \pi]$  is the angular distance of the rotational part between the estimated,  $\mathbf{P}_{t,l}^l$ , and the ground-truth hand pose at time  $t$  for trajectory  $l$ ,  $\hat{\mathbf{P}}_{t,l}^l$ .

For the end-effector, we compute the score  $\sigma_3(\mathcal{E}, \hat{\mathcal{E}}, \varepsilon)$  using Eq. 4 without summation over time. The value of the threshold  $\varepsilon$  is set to 3 cm for both measures.

The final benchmark score,  $S \in [0, 1]$ , is

$$S = \frac{1}{3} \sum_{j=1}^{13} \lambda_j s_j, \quad (5)$$

where  $\lambda_j = \frac{1}{9}$  for  $j \in \{1, 2, 3\}$  (width at top, width at bottom and height);  $\lambda_j = \frac{1}{3}$  for  $j \in \{4, \dots, 10\}$ ;  $\lambda_j = \frac{1}{12}$  for  $j \in \{11, 13\}$  (human maneuvering time and robot maneuvering time); and  $\lambda_{12} = \frac{1}{6}$  (handover time).

#### IV. THE BASELINE

##### A. Overview

To facilitate participation in the benchmark, we share the code for a baseline as open source<sup>5</sup>, and instantiate this baseline in two setups, S1 and S2, in different laboratories. The baseline estimates the centroid and the dimensions of the cup in 3D at 18 Hz on an Intel i7-7700 CPU @ 3.60GHz, 16GB RAM, and a GeForce GTX 1060 6GB GPU.

The setup in S1 uses a KUKA LBR iiwa 7-DoF manipulator (14 kg payload)<sup>6</sup> and a task-space control of the robotic manipulator, and prediction/inference of the human intention. The setup in S2 uses a UR5 6-DoF manipulator (5 kg payload)<sup>7</sup>

and a simpler robotic control with standard motion planning libraries<sup>8</sup>. Both robots are equipped with a Robotiq 2F-85 2-finger gripper<sup>9</sup>.

Both setups use two cameras (Intel RealSense D435), located at 40 cm from the arm. The cameras view the center of the table and record RGB sequences at 30 Hz with a resolution of 1280×720. The cameras are synchronized, calibrated and localized with respect to a calibration board. Fig. 4(a) shows an example of the RGB images captured from setup S2.

##### B. Vision pipeline

We combine semantic segmentation with object tracking in a two-stage algorithm that first estimates the 3D object location using multi-view projective geometry [12] and then estimates the height and width of the object with a generative 3D sampling model. We assume that only the semantic category of the object (i.e. a drinking cup) is known, and that the object is symmetric with respect to its major axis and initially is on the table.

Let  $I_t^c \in \{0, 255\}^{W,H,C}$  be the image acquired at time  $t$  by camera  $c \in \{0, 1\}$ , where  $W, H, C$  are the width, height, and number of channels of the image, respectively. Let  $\mathbf{O}_t = (x, y, z, w, h)$  represent the cup in 3D, where  $\mathbf{O}'_0 = (x, y, z)$  is its centroid,  $w$  the width at the centroid, and  $h$  the height.

Let an object segmentation algorithm take an image,  $I_t^c$ , as input and return a binary mask,  $\mathbf{m}_t^c$ , where the non-zero pixels denote the location of the object on the image plane. We localize the object in the first frame,  $I_0^c$ , of each camera using MaskRCNN [35]. We trained this algorithm with images of generic cups and plastic wine glasses from the COCO dataset [36] with a ResNet-50-FPN backbone [37][38], pre-trained on ImageNet [39]. In the validation dataset, the semantic segmentation algorithm achieves 67.6% and 73.1% average precision (intersection over union based overlap ratio greater than 50%) for the detection and segmentation task, respectively.

To estimate the 2D centroid of the segmented object from the mask,  $\mathbf{C}_0^c$ , in the first frame we use the intensity centroid method [40]. The 2D centroid from each camera is then used in a triangulation process to locate the centroid in 3D [12]:  $\mathbf{O}'_0 = \gamma(\mathbf{C}_0^1, \mathbf{C}_0^2, \Theta^1, \Theta^2)$ , where  $\gamma(\cdot)$  is the triangulation function, and  $\Theta^1$  and  $\Theta^2$  are the projection parameters of Camera 1 and Camera 2, respectively. The projection parameters include focal length and principal point (the intrinsics), and the location and orientation of the camera in 3D (camera pose or extrinsics) with respect to the reference coordinate system, which is related to a calibration checkerboard. Fig. 4(c) shows an example of the estimated dimensions and 3D location of the cup.

To determine the dimensions of the cup, we first estimate its height as  $h = 2z$ . To estimate its width, we use a 3D generative model that samples 3D points from multiple concentric circumferences centered at the height of the centroid and parallel to the surface of the table. The diameter of the largest circumference is chosen based on the maximum aperture of

<sup>5</sup>Code available at <https://github.com/CORSMAL/Benchmark/>

<sup>6</sup><https://www.kuka.com/en-ch/products/robotics-systems/industrial-robots/lbr-iiwa/>

<sup>7</sup><https://www.universal-robots.com/products/ur5-robot/>

<sup>8</sup>MoveIt <https://moveit.ros.org/>

<sup>9</sup><https://robotiq.com/products/2f85-140-adaptive-robot-gripper/>

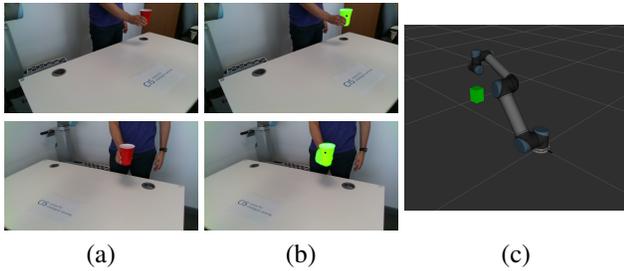


Fig. 4. Sample segmentation and tracking result: (a) input frames (top: Camera 1, bottom: Camera 2); (b) 2D tracking result (green mask) and re-projected centroid (blue point); (c) visualization of the 3D location and dimensions of the cup (green) estimated through vision.

the gripper (8.5 cm in setups S1 and S2) and the diameter of the smallest circumference is 5 cm.

We then project each 3D point  $\mathbf{X}_n$  into the image plane of each camera as  $\mathbf{u}_n^c = \pi(\mathbf{X}_n, \Theta^c)$ , where  $\pi(\cdot) : \mathbb{R}^3 \rightarrow \mathbb{R}^2$  is the projection function [12]. To find the optimal width, we verify if all points belonging to each circumference, from the largest to the smallest, lie within  $\mathbf{m}_0^c$ , the segmentation mask of the object. The first circumference that satisfies the condition approximates the width of the object of interest at the height of the estimated centroid. As the estimation is independent for each camera, the final width is computed as their average. To estimate the width at the base of the cup,  $w_b^i$ , we repeat this process by generating the 3D circumference at  $z = 0$ . Finally, we estimate  $w^i$ , the width at the top of the cup, via linear extrapolation from the knowledge of the width at the base and the width at the centroid.

To deal with changes in appearance and to produce a smooth trajectory of the centroid for the controller of the robotic subsystem, we track the cup over time [41]. The tracker,  $\mathcal{T}(\cdot)$ , operates in each camera independently and is automatically initialized with the smallest rectangle containing all pixels of the mask  $\mathbf{m}_0^c$ . The tracker produces for each frame a binary mask for the object,  $\mathbf{m}_t^c = \mathcal{T}(\mathbf{m}_{t-1}^c, I_t^c)$ , which we use to estimate the centroid of the cup in 3D with the same procedure employed for the segmentation masks. Fig. 4(b) shows two sample masks produced by the tracker for Camera 1 and Camera 2 in S2.

### C. Robotic control

The robotic control uses  $\mathbf{O}_t$ , the centroid and dimensions of the cup estimated through vision, to grasp the cup, whose pose during the handover shall be upright. We approximate the location of the hand that handles the cup with the estimated centroid of the cup. We set the desired end-effector pose at time  $t$  as  $\mathbf{O}'_t$ , with constant value for the orientation and we use the estimated width,  $w$ , to accurately close the gripper once the end-effector reaches the vicinity of the cup.

We control the robot in the end-effector (or task) space. Similarly to [42], we formulate the motion and constraint equations for a robotic manipulator as a Quadratic Programming (QP) problem; when the robot is torque-controlled we use the formulation for *inverse dynamics*, otherwise we use the

formulation for *inverse kinematics*. We compute the desired end-effector accelerations or velocities to pass to the QP problem by using a PD controller in task-space based on our algorithm that predicts the next end-effector target based on the movement of the cup.

For the handover to be effective, the robotic manipulator needs to predict the motion of the human hand and try to intercept it. We use an approach inspired by human-to-human handovers [43]. For each handover  $l$ , the distance between the giver-receiver hands,  $\xi_l$ , is modelled as Gaussian Mixture Model (GMM), whose parameters  $\theta$  (prior probability, mean value, and covariance) are learned via Expectation Maximization [44] using a dataset of human-to-human handover trajectories,  $\mathcal{P}$ , for both the giver and the receiver. The learned GMM parameters are then encoded in a Coupled Dynamical System,  $\mathcal{P}(\Psi(\xi_l), \xi_l | \theta)$ , which defines a coupling function,  $\Psi(\xi_l)$ , between the giver-receiver hands during handovers:  $\Psi(|d_h|) = |d_p|$ , where  $d_h$  and  $d_p$  represent the distance  $\xi_l$  in the z-axis (orthogonal to the table) and x-axis (tangential with respect to the table) of the giver-receiver hands, respectively.

For the handover, we command the end-effector to follow the path a human receiver would follow. In particular, the controller uses the current distances  $d_h$  and  $d_p$  between the human hand and the robot end-effector, and outputs the predicted next distance at which the robot should be from the human hand. We then integrate forward in time to get the desired acceleration (or velocity) for the end-effector to follow. We continue this tracking-inference cycle until the distance between the current and the desired pose of the end-effector of the manipulator is smaller than a threshold, which depends on the estimated width of the object<sup>10</sup>, and then perform the handover by closing the gripper.

Note that because the mass of the cup is unknown, the controller can produce motions that spill the content. Thus, we encourage participants to develop techniques (or sensors) that estimate the payload. For our baseline, we ignored the estimation and relied on the task-space PD-gains. Fig. 5 shows one unsuccessful and two successful handovers with different grasp types and handover locations.

### D. Discussion

The last two columns of Table II report the results on the benchmark for the baseline in setups S1 and S2<sup>11</sup>. The overall benchmark score is 0.32 for S1 and 0.33 for S2. The vision score is 0.19 for S1 and 0.20 for S2. The robot score is 0.31 for both S1 and S2, and the task score is 0.46 for S1 and 0.40 for S2. The delivery location score is 0.47 for S1 and 0.37 for S2, thus reflecting the more advanced controller used in S1; whereas the score for the mass of the delivered filling is 0.49 for S1 and 0.58 for S2, which somehow is a reflection of the speed of the robot maneuvers (0.45 for S1 and 0.09 for S2). Note that the scores also include failure cases and

<sup>10</sup>We used the estimated width as the threshold in our experiments.

<sup>11</sup>Note that the scores for  $s_4$ ,  $s_5$ ,  $s_6$  and  $s_7$  are set to null as the baseline does not estimate the mass and fullness of the cups. Advanced algorithms should be developed to estimate these properties for manipulation by observing the object itself and/or the maneuvering of the object by the human.

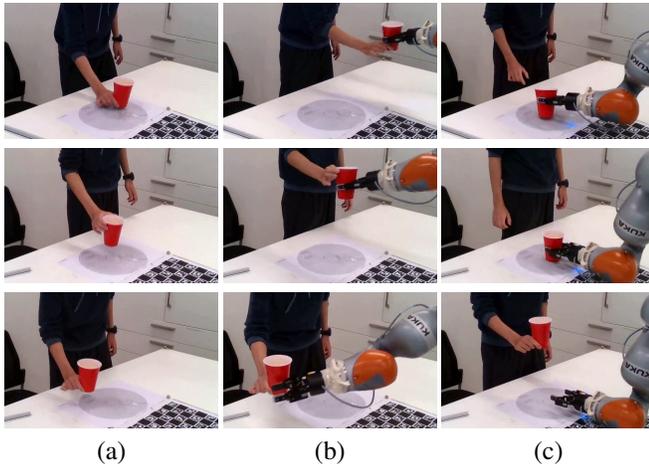


Fig. 5. Sample handovers from setup S1, viewed from Camera 1, for Cup 2, full with Grasp 1 (first row), with Grasp 2 (second row), and with Grasp 1 (third row). (a) Human manipulation. (b) Handover. (c) Robot manipulation (third row: a failed handover).

unstable tracking for some configurations (e.g. cups with high transparency: Cup 3 and Cup 4). Interestingly, from the raw measures, the handover time was on average about a second in both setups, while the typical human-to-human handover time is less than 300 ms [26], suggesting a clear direction for improvement.

Fig. 6 shows aggregated scores by cup, filling, and grasp type as well as handover location. Unlike Grasp 1 and Grasp 2, the baseline does not allow the performance of safe handovers when the subject holds the cup with a natural grasp (Grasp 3). This is due to the large size of the gripper we use: advanced perception algorithms should estimate the pose of the cup and human hand to detect suitable grasp regions to inform grasp planning. Note also that the performance is similar for the three handover locations, but varies for different cups, with the plastic wine glass (Cup 4) being the most challenging.

Overall, the scores represent baseline results that we expect the research community to improve through new perception and control algorithms that aim to perform seamless human-to-robot handovers.

## V. CONCLUSION

We proposed a benchmark to evaluate dynamic human-to-robot handovers in scenarios without motion capture systems, markers, or prior object models. We consider previously unseen objects (drinking cups), whose physical properties are subject to transformations, such as deformability due to the grasp, or different stiffness and filling amounts. Along with the benchmark, we presented a baseline implementation and released its code, which we hope will encourage the community to participate in this benchmarking effort.

As future work, we will extend the benchmark with new objects and fillings for the inference of further physical properties. Also, we will include in our baseline tactile sensing to improve the perception capabilities of the robot with the goal of achieving human-level flexibility and adaptability in grasping previously unseen objects handed over by a person.

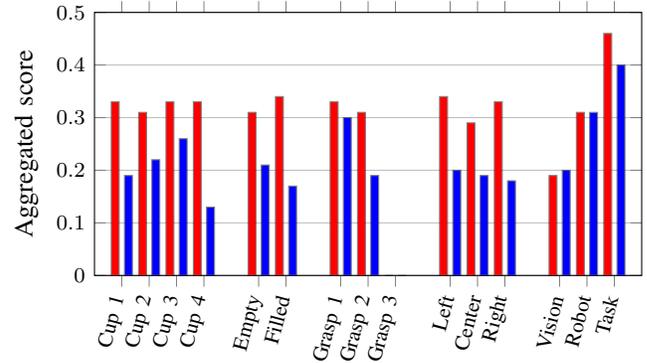


Fig. 6. Scores averaged by cup, fillings, and grasp type; handover location; and subsystem for setup S1 (■) and setup S2 (■). Note that the baseline does not cope with handovers with Grasp 3, when most of the surface of the cup is occluded by the hand of the subject.

## REFERENCES

- [1] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "PVNet: Pixel-wise voting network for 6DOF pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 16–20 June 2019.
- [2] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 16–20 June 2019.
- [3] A. Collet, M. Martinez, and S. S. Srinivasa, "The MOPED framework: Object recognition and pose estimation for manipulation," *Int. J. Robot. Res.*, vol. 30, no. 10, pp. 1284–1306, 2011.
- [4] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Proc. Robotics: Science and Syst.*, Pittsburgh, USA, 26–30 June 2018.
- [5] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, and D. Fox, "PoseRBPF: A Rao-Blackwellized Particle Filter for 6D Object Pose Tracking," in *Robotics: Science and systems workshop*, Freiburg im Breisgau, Germany, 22–26 June 2019.
- [6] J. R. Medina, F. Duvallet, M. Karnam, and A. Billard, "A human-inspired controller for fluid human-robot handovers," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, Cancun, Mexico, 15–17 Nov. 2016.
- [7] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set," *IEEE Robotics Autom. Magazine*, vol. 22, no. 3, pp. 36–52, 2015.
- [8] G. Koostra, M. Popović, J. A. Jørgensen, D. Kragic, H. G. Petersen, and N. Krüger, "VisGraB: A benchmark for vision-based grasping," *J. Behavioural Robotics*, vol. 3, no. 2, pp. 54–62, 2012.
- [9] T. Wang, C. Yang, F. Kirchner, P. Du, F. Sun, and B. Fang, "Multimodal grasp data set: A novel visual-tactile data set for robotic manipulation," *Int. J. Advanced Robotic Syst.*, vol. 16, no. 1, pp. 1–10, 2019.
- [10] A. H. Mason and C. L. MacKenzie, "Grip forces when passing an object to a partner," *Experimental Brain Research*, vol. 163, no. 2, pp. 173–187, 2005.
- [11] S. Kim, A. Shukla, and A. Billard, "Catching objects in flight," *IEEE Trans. Robotics*, vol. 30, no. 5, pp. 1049–1065, 2014.
- [12] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2nd ed. Cambridge University Press, 2003.
- [13] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "Deepim: Deep iterative matching for 6D pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 8–14 Sept. 2018.
- [14] Y. Hu, J. Hugonot, P. Fua, and M. Salzmann, "Segmentation-driven 6D object pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 16–20 June 2019.
- [15] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robot. Res.*, vol. 27, no. 2, pp. 157–173, 2008.

- [16] A. Saxena, L. Wong, M. Quigley, and A. Y. Ng, "A vision-based system for grasping novel objects in cluttered environments," in *Proc. Int. Symp. Robot. Res.*, Hiroshima, Japan, 26–29 Nov. 2007.
- [17] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, no. 4-5, pp. 705–724, 2015.
- [18] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. Abbeel, "Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding," in *Proc. IEEE Int. Conf. Robotics Autom.*, Anchorage, AK, USA, 3–8 May 2010.
- [19] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *Proc. IEEE Int. Conf. Robotics Autom.*, Stockholm, Sweden, 16–21 May 2016.
- [20] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1577–1591, 2013.
- [21] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio-visual sensing device," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2576–2588, 2019.
- [22] I. Skrypnik and D. G. Lowe, "Scene modelling, recognition and tracking with invariant image features," in *IEEE/ACM Int. Symp. Mixed Augmented Reality*, Arlington, VA, USA, 5 Nov. 2004.
- [23] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPhP: An accurate O(n) solution to the PnP problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.
- [24] V. Lepetit and P. Fua, "Monocular model-based 3D tracking of rigid objects: A survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 1, no. 1, pp. 1–89, 2005.
- [25] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6D object pose and size estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 16–20 June 2019.
- [26] E. C. Grigore, K. Eder, A. G. Pipe, C. Melhuish, and U. Leonards, "Joint action understanding improves robot-to-human object handover," in *Proc. IEEE Int. Conf. Intell. Robot Syst.*, Tokyo, Japan, 3–8 Nov. 2013.
- [27] W. P. Chan, C. A. Parker, H. M. Van der Loos, and E. A. Croft, "A human-inspired object handover controller," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 971–983, 2013.
- [28] W. P. Chan, M. K. Pan, E. A. Croft, and M. Inaba, "Characterization of handover orientations used by humans for efficient robot to human handovers," in *Proc. IEEE Int. Conf. Intell. Robot Syst.*, Hamburg, Germany, 28 Sept./2 Oct. 2015.
- [29] J. Waldhart, M. Gharbi, and R. Alami, "Planning handovers involving humans and robots in constrained environment," in *Proc. IEEE Int. Conf. Intell. Robot Syst.*, Hamburg, Germany, 28 Sept./2 Oct. 2015.
- [30] J. Mainprice, M. Gharbi, T. Siméon, and R. Alami, "Sharing effort in planning human-robot handover tasks," in *Proc. IEEE Int. Symp. on Robot and Human Interact. Comm.*, Paris, France, 9–13 Sept. 2012.
- [31] M. Cakmak, S. S. Srinivasa, M. K. Lee, J. Forlizzi, and S. Kiesler, "Human preferences for robot-human hand-over configurations," in *Proc. IEEE Int. Conf. Intell. Robot Syst.*, San Francisco, CA, USA, 25–30 Sept. 2011.
- [32] V. Micelli, K. Strabala, and S. Srinivasa, "Perception and control challenges for effective human-robot handoffs," in *Robotics: Science and systems workshop*, Los Angeles, CA, USA, 27–30 June 2011.
- [33] M. Prada, A. Remazeilles, A. Koene, and S. Endo, "Dynamic movement primitives for human-robot interaction: comparison with human behavioral observation," in *Proc. IEEE Int. Conf. Intell. Robot Syst.*, Tokyo, Japan, 3–7 Nov. 2013.
- [34] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *Int. J. Robot. Res.*, vol. 36, no. 3, pp. 261–268, 2017.
- [35] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 22–29 Oct. 2017.
- [36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 8–14 Sept. 2018.
- [37] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 21–26 July 2017.
- [38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 27–30 June 2016.
- [39] J. Deng, W. Dong, R. Socher, L.-J. Li, L. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 20–25 June 2009.
- [40] P. L. Rosin, "Measuring corner properties," *Comput. Vis. Image Understanding*, vol. 73, no. 2, pp. 291–307, 1999.
- [41] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 16–20 June 2019.
- [42] S. Feng, E. Whitman, X. Xinjilefu, and C. G. Atkeson, "Optimization-based full body control for the darpa robotics challenge," *J. of Field Robotics*, vol. 32, no. 2, pp. 293–312, 2015.
- [43] N. Ferreira Duarte, M. Raković, and J. Santos-Victor, "Coupling of arm movements during human-robot interaction: the handover case," in *Proc. IEEE Int. Symp. on Robot and Human Interact. Comm.*, New Delhi, India, 14–18 Oct. 2019.
- [44] C. M. Bishop, *Pattern recognition and machine learning*, 2nd ed. Springer, 2006.