2023 32nd IEEE International Conference on Robot and
Human Interactive Communication (RO-MAN)
August 28-31, 2023. Paradise Hotel, Busan, Korea

2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) | 979-8-3503-3670-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/RO-MAN57019.2023.10309337

Depth image-based deformation estimation of deformable objects for collaborative mobile transportation

Giorgio Nicola[1], Stefano Mutti[1], Enrico Villagrossi[1], Nicola Pedrocchi[1]

***Abstract*— Human-Robot collaborative transportation is a promising technology that combines the strength of humans and robots. The most common approaches rely on methodologies that exploit force-sensing. However, the drawbacks are multiple. First, the magnitude of force applied might be limited to avoid damages. Then, force measurements might be unidirectional according to the material properties; e.g., compression forces are not measurable for fabrics. This paper proposes an approach based on the estimation of the deformation state of the manipulated object from depth images. Specifically, the segmented depth image of the manipulated object are fed to a Convolutional Neural Network (CNN) model to estimate the current deformation status. Compared with the desired deformation, the current deformation status is used to generate the robot's twist command. The methodology is proved in a mobile robot application, where carbon-fiber fabrics are transported. A comparison with the state-of-the-art is reported proving that the proposed method is more accurate and more repeatable.**

I. Introduction

Human-robot collaborative transportation is increasingly investigated for industrial applications, mainly when applied to large objects requiring multiple people to handle. However, multiple challenges need to be solved still. First, the robot should be able to infer the human objective [1], and based on that, the human and robot should be able to switch the leader and follower roles during the physical interaction [2]. Second, the robot should minimize human effort while ensuring safety [3]. Then, human control action in physical human-robot interaction can be approximated as a delayed linear model because of the human reaction time [4]. Such a reaction time varies significantly between people and, if not addressed, can even lead to instability.

A peculiar case of human-robot collaborative transport is when the manipulated object is deformable, such as fabric or cables. The robot should avoid excessive deformation of the manipulated material and aid the human simultaneously. Standard control techniques, based on sensing the force applied by the human, e.g., impedance control, struggle due to the low forces that can be applied before damaging the object. Thus, control strategies based on visual feedback of the comanipulated object are viable alternatives.

Building on our previous work [5], we propose a vision-based strategy to perform collaborative transportation of deformable objects with an industrial mobile manipulator (IMM). We use an eye-in-hand RGB-D camera to capture the depth image of the manipulated object. The segmented image of the manipulated object is fed to a Convolutional Neural Network (CNN) to estimate its current deformation status, defined as the current human-robot relative distance. The difference between the current deformation status and a predefined rest configuration status is finally converted into a twist command and fed to the robotic controller.

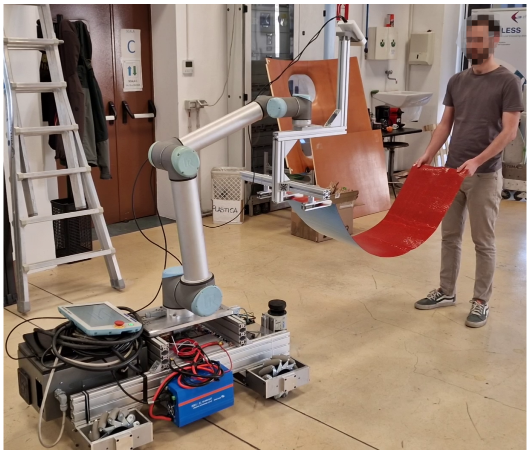

Fig. 1: Developed setup for collaborative transportation of deformable objects. A sheet of carbon fiber fabric is manipulated by a human and a mobile robot. An RGB-D camera is used to sense the deformation of the carbon fiber sheet.

A. Related Work

The manipulation of deformable objects has been investigated [6], [7] focusing on two main classes of deformable objects: cables [8], [9] and cloth-like [10], [11]. However, a few works study the problem of the collaborative human-robot manipulation of such objects, and they typically focus on cloth folding [12], [13]. One of the main challenges is tracking the deformable object and estimating its current shape, in other words, its deformation. Two main approaches have been developed to estimate the material deformation in human-robot manipulation: direct or indirect via motion capture of the human. The direct approach uses visual features that are subsequently converted into robot commands. However, multiple visual features have been developed in the literature; therefore, a standardized control architecture has yet to be established. Indeed in [14], fabric folds combined with force measurements are used. In [15], [16], the authors use Histograms of Oriented Wrinkles (HOWs) computed by applying Gabor filters to RGB images. The indirect approach instead tracks the position of the human grasping points on

Giorgio Nicola *et al.* are with the Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing, National Research Council of Italy, Via A. Corti 12, 20133, Milan, Italy; *giorgio.nicola@stiima.cnr.it*

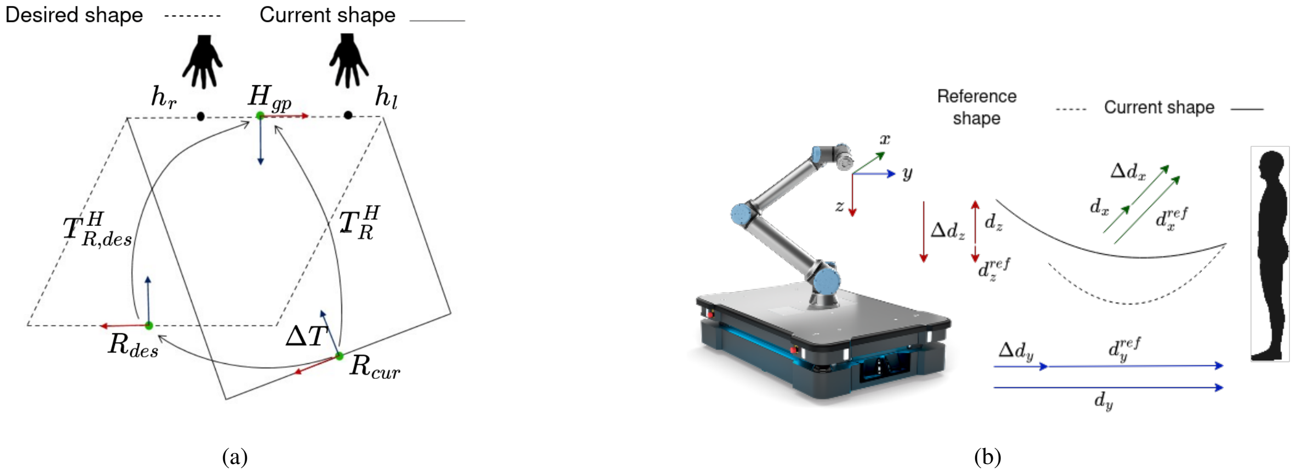

Fig. 2: Problem formulation of the human-robot collaborative transportation problem. a) Shows the top view highlighting the definition of the human grasping point $\mathbf{H}_{gp}$. b) Shows a lateral view highlighting the parameters that compose $\mathbf{T}^{H}_{R}$, $\mathbf{T}^{H}_{R,des}$, and $\Delta\mathbf{T}$. For the sake of simplicity, rotations have been neglected.

the material via a motion capture system based either on IMU sensors [17], combined with force/torque measurement for stiffer materials, or a camera [18], [19]. However, the general assumption is that the contact point between the human and the manipulated material is known *a priori* or detectable.

A different approach is in [20]. A motion tracker detects handcrafted coded gestures that, combined with torque force measures, are used directly to compute robot end effector speed without estimating the material deformation.

Deep Neural Networks have recently been used as feature extractors via autoencoder networks [21], [22], [23], and the robot movement is planned according to such features. However, those methods have yet to be applied to human-robot manipulation.

Finally, only a few works study the human-robot collaborative transportation with mobile manipulators [20], [17], while many works investigate non-collaborative scenarios with multiple mobile manipulators [24], [25].

Collaborative transportation has been investigated mostly applied to rigid objects, such as in [26], [27]. The general approach is based on force sensing combined with impedance or admittance control. However, force measurements with deformable materials like fabrics are unreliable due to i) low forces sustained by the object before damage; ii) force measurement might be unidirectional, specifically only traction forces; iii) translation-rotation ambiguity. Thus such control architecture cannot be applied to deformable objects without the aid of vision-based deformation estimation such as in [17].

B. Contribution

The approach proposed in this paper relies on learning a deformation model of the manipulated objects from depth images and converting it into robot twist commands. Compared with methods in the literature, it has multiple advantages. Firstly, the deformation estimation is direct yet not based on handcrafted visual features; instead, features are learned and can be general with a sufficiently large dataset. Furthermore, indirect deformation estimation based on a motion capture system has various drawbacks, with IMU-based skeleton tracking being affected by drift, body sensors needing to be worn by the operator with uncomfortable straps or specific suits, and camera-based skeleton tracking having a limited field of vision. Vision-based methods are also generally computationally expensive, requiring high-end hardware to reach 30 Hz (standard RGB-D cameras' frame rate upper limit). Second, the proposed control architecture is very straightforward compared to other works using a combination of force and vision-based control architectures.

II. Method

A. Problem formulation

The problem of human-robot collaborative manipulation of deformable materials is composed of two agents handling the deformable material simultaneously, as shown in Figure 1. One agent is the human that leads the activity, and the second one is the robot that should follow the human movement. The objective is to manipulate the desired object while minimizing the deformations from a rest configuration that guarantees no damage to the material. The human movements during the collaborative transport deform the material; thus, the robot compensating for the deformations follows the human.

The problem of collaborative transportation can be decomposed into two separate problems. First is the definition and estimation of the deformation status. Second is computing the robot commands to minimize the deformation from the rest desired position.

We briefly report the definition of deformation status first proposed in [5] and expanded in [28]. Let us consider a deformable material handled by two agents, a human, and a robot. Denote h_r and h_l as the human's right and left hand grasping positions and R_{cur} as the robot's current grasping point. The object shape can be described as the tuple of relative roto-translations between the robot grasping

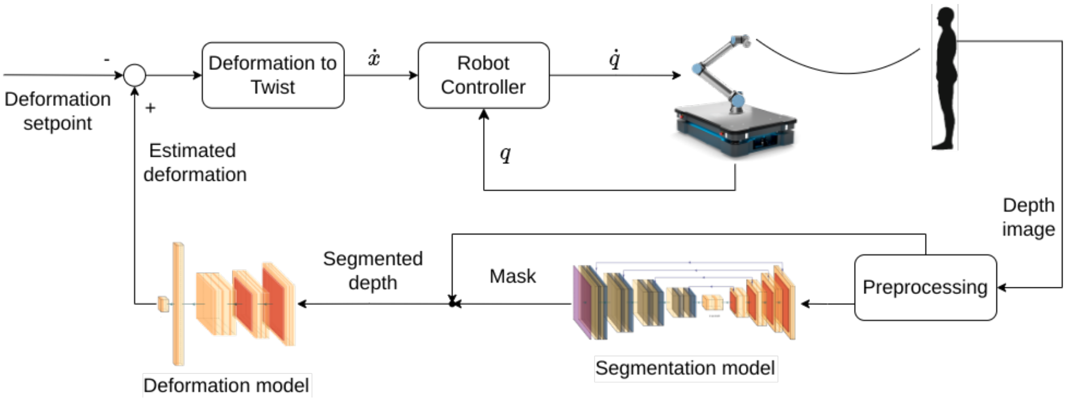

Fig. 3: Architecture for the collaborative transportation problem. After preprocessing, the depth image acquired is segmented to obtain the segmented depth image of the comanipulated object. The segmented depth is fed to a deformation model that estimates the current deformation status as defined in Section II-A. The current delta deformation between the estimated and desired setpoint is converted into a twist command fed to the robot controllers that output the IMM joints' speed.

position and the human holding positions, except for the local deformation around the hands. This model fully describes the problem of collaborative shape servoing but is unnecessarily accurate for the objective of collaborative transportation. It also models the deformation in the corners depending on where the human grasps the object.

We also assume that the human does not deform the object between the hands, considering the material segment between the two hand-grasping points as rigid, for example, spreading the arms. Such movements do not have a purpose in a collaborative transportation application; therefore, we assume that if performed are involuntary and neglectable. Given these assumptions, we define an arbitrary point on the object between the hands invariant to the different grasping positions. The relative roto-translations between the robot grasping position and this point is a robust proxy for the object deformation, except for the local deformation around the hands and the geometry of the corners, which are irrelevant in a collaborative transportation scenario. Finally, given the frame reference in Figure 2b, we also assume that the x-axis rotation is always zero since it causes only local deformations of the object around the robot grasping point.

Referring to Figure 2a, $\mathbf{H}_{gp}$ is the human grasping point frame, $\mathbf{T}_R^H$ and $\mathbf{T}_{R,des}^H$ are the actual roto-translation matrix between the robot's current pose $\mathbf{R}_{cur}$ or desired $\mathbf{R}_{des}$ and $\mathbf{H}_{gp}$. $\mathbf{T}_R^H$ and $\mathbf{T}_{R,des}^H$ are respectively described by the parameters d_R^H and $d_{R,des}^H$, three translations and three rotations following the X-Y-Z extrinsic Euler conventions.

Given this formulation, the problem consists of (i) imposing the target $\mathbf{T}_{R,des}^H$ and (ii) estimating $\mathbf{T}_R^H$ the collaborative transportation. The robot should be controlled based on such estimations to minimize the distance from the target's desired pose, i.e., $\Delta T \rightarrow I$, where I is the identity matrix. Given the formulation above, we define the robot controller as follows:

$$\dot{q} = f(\Delta\mathbf{T}) \quad \Delta\mathbf{T} \rightarrow I \tag{1}$$

where $\dot{q}$ denotes the robot joint command speed.

B. Proposed solution

We solve the collaborative transportation with the architecture shown in Figure 3. An RGB-D camera is mounted on the robot end effector to acquire depth images of the manipulated object; see Figure 1. The background and the human partner in the depth images are segmented with a specifically trained segmentation model, implemented using Unet [29] with the encoder-decoder pre-trained on the ImageNet dataset. The segmented depth image of the object is then fed to the deformation model, based on DenseNet121 [30] pretrained on ImageNet, which estimates roto-translation $\mathbf{T}_R^H$ parameters. Both the segmentation and deformation models are trained on a single dataset specifically acquired of deformed plies. Subsequently, the delta deformation between the deformation setpoint $\mathbf{T}_{R,des}^H$ and the estimated deformation $\mathbf{T}_R^H$ is converted into a twist command which is then handled by the robot control algorithm.

C. Datasets acquisition

As described above, we developed a procedure to acquire a single dataset to train the deformation and segmentation models. The dataset acquisition is divided into two steps. First, the deformed object dataset is acquired, called the deformation dataset. Second, the dataset for the segmentation model, called the segmentation dataset, is generated starting from the deformation dataset. This approach allows for minimizing the required time to acquire datasets and minimizing the necessary human intervention as much as possible.

The deformation dataset should comprise many depth images of the deformed material with different human-robot relative distances and human grasping positions. We substituted the human with a frame that holds the deformable material, as in Figure 4, to avoid inaccuracies introduced by a human operator. Then rather than deforming the object moving the human or frame, the robot moves relative to the frame whose position is estimated through a pair of fiducial markers AprilTags. At the same time, the frame allows higher accuracy and repeatability, simulating different human

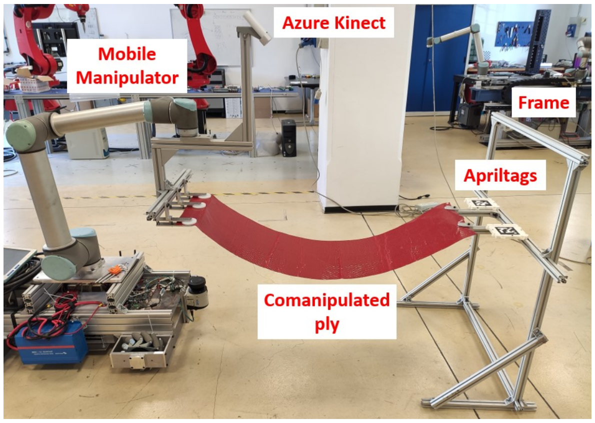

Fig. 4: The setup to acquire the dataset of a carbon fiber object. It comprises an RGB-D camera, Azure Kinect, an IMM composed of a Universal Robot UR10 and a custom-made mobile platform, an aluminum frame to mimic the human, and a pair of fiducial markers, Apriltags, to localize the frame.

grasping positions on the material, and minimizes human intervention. Finally, at each robot position, a set of RGB-D images are taken to account for the camera's noisy output and are autonomously labeled with T_R^H parameters described in Section II-A. The deformation dataset acquisition is almost fully autonomous since the only human intervention required is to modify the grasping position on the frame.

Finally, the deformation estimation model should be agnostic to anything from the object shape, such as the background or human partner. Thus, the depth images are segmented with a fully autonomous segmentation routine. The segmentation routine depicted in Figure 5 uses thresholding of the depth, static mask, and a mask based on Apriltags' position in the RGB image. The Apriltags define a rectangular box in the depth image, and all pixels outside are set to zero. Each image is cropped and resized to the resolution of 224×224. During the deformation model training, random translation and random rotations are applied to the segmented depth image as a form of data augmentation.

Such a segmentation routine cannot be unlikely applied in collaborative transportation since the operator must wear a pair of fiducial markers. Therefore, the segmentation model should be able to segment the object independently by the background and the human operator. Thus, instead of acquiring and labeling a new dataset sufficiently large to generalize over different backgrounds and humans, we generated a synthetic dataset starting from the segmented depth images of the deformation dataset. In detail, the synthetic dataset is created by combining a minimal dataset of human-depth images, 100 images per 5 random people, from the same point of view, in random positions not overlapping with the segmented object, as shown in Figure 5. Finally, fake backgrounds with the depth value thresholded are added.

D. Mobile manipulator control

Consider (1). The mobile manipulator controller should compute robot joints speed $\dot{q}$ taking as input the deformation error, *i.e.*, ΔT. First, we compute twist commands from deformation, as in Figure 3, then we convert such commands into joint speed commands.

The deformation module outputs the parameters describing $\mathbf{T}_R^H$ denoted as $d_R^H \in \mathbb{R}^6$, as shown in Section II-B. Following the same rotation convention, we compute $\mathbf{T}_{R,des}^H$ parameters, $d_{R,des}^H$. The delta deformation Δd turns in:

$$\Delta d = d_R^H - d_{R,des}^H . \tag{2}$$

Twist command, denoted as $\dot{x}_{tool}$, is computed with the same conventions with a simple gain $K \in \mathbb{R}^6$.

$$\dot{x}_{tool} = K \odot \Delta d \tag{3}$$

Based on the formulation in Section II-A, ΔT is defined in the frame $\mathbf{R}_{cur}$, therefore also $\dot{x}_{tool}$, which is then converted to the IMM base frame H_{mr}, and referred to as $\dot{x}$ for the sake of simplicity.

Considering an n degrees of freedom IMM, with joint velocity $\dot{q} \in \mathbb{R}^n$, formed by the robotic arm joint velocities and by the generalized velocities of the AMR (Autonomous Mobile Robot), and task space velocity $\dot{x} \in \mathbb{R}^m$, their nominal relation follows

$$\dot{q} = \mathbf{J}(\mathbf{q})^+ \dot{x} \tag{4}$$

where $\mathbf{J}(\mathbf{q})^+ \in \mathbb{R}^{n*m}$ is the pseudo-inverse of the task Jacobian matrix referred to the IMM base $\mathbf{H}_{mr}$.

The task Jacobian $\mathbf{J}(\mathbf{q})$ is formed by the robotic arm Jacobian $\mathbf{J}(\mathbf{q})_\mathbf{R}$, referred to the H_{mr} frame, followed by the trivial AMR Jacobian as

$$\begin{bmatrix} \mathbf{J}(\mathbf{q})_\mathbf{R} & \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{matrix} \end{bmatrix} \tag{5}$$

Being the system redundancy ($m < n$), we enforce additional constraints to keep the robotic arm as close as possible to a resting given position q_{rest} during the control, as:

$$\dot{q}_{err} = K_p(q_{rest} - q) \tag{6}$$

$$\dot{q} = \mathbf{J}(\mathbf{q})^+ \dot{x} + (I - \mathbf{J}(\mathbf{q})^+ \mathbf{J}(\mathbf{q})) \dot{q}_{err} \tag{7}$$

and K_p is a proportional gain that affects the null movement's magnitude. The AMR compensates for displacements and aids the robotic arm in keeping joint positions close to the q_{rest}, preventing the reaching of joint limits.

III. Experiments and Results

A. Experimental setup

The experimental setup shown in Figure 1 comprises a Universal Robot UR10 mounted on top of a custom-made

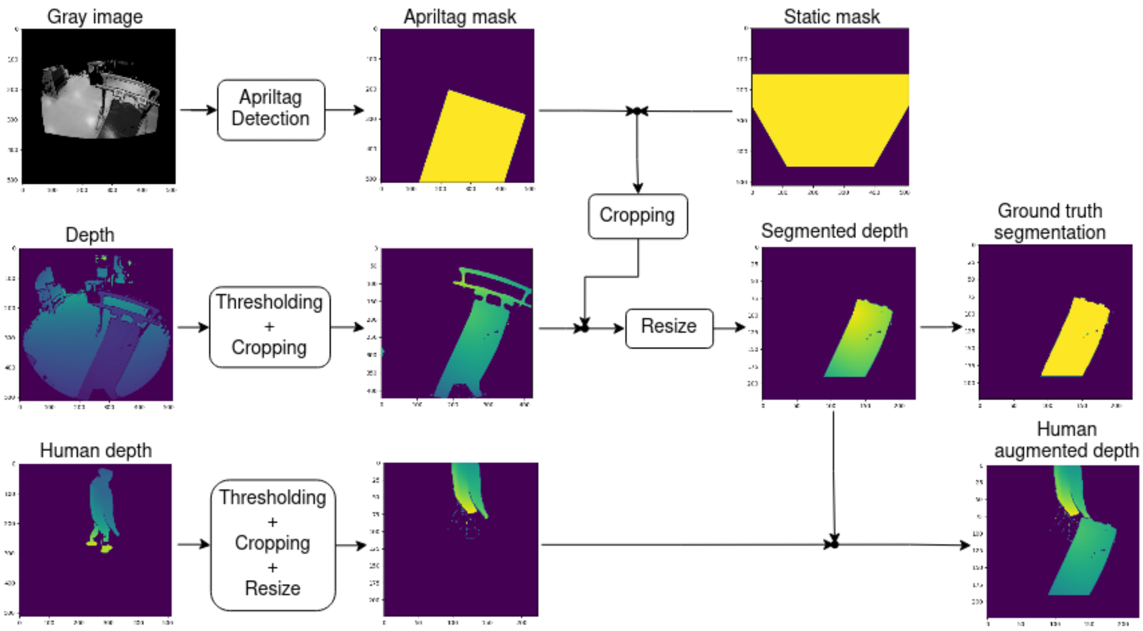

Fig. 5: Developed pipeline to generate data for training the segmentation and deformation estimation models. The segmented depth is generated by combining masks, thresholding, and cropping. The segmented depth images are combined with humans' depth images to train the segmentation model to segment the object from humans not included in the dataset.

omnidirectional mobile platform and an Azure Kinect RGB-D camera. A roughly rectangular carbon fiber ply with size 130×38 $[cm]$ was used. The acquired deformation dataset comprises 134550 depth images of the ply deformed along the three translations and the rotation along the z-axis. Five images are obtained for each robot pose, i.e., human-robot relative distance, and the dataset includes two different human grasping positions. Given the desired resting deformation of $[0,\ 1.05, 0, 0]$ the studied deformation range was:

- $-0.12 \leq x \leq 0.12$ $[m]$,
- $-0.255 \leq y \leq 0.255$ $[m]$,
- $-0.12 \leq z \leq 0.12$ $[m]$,
- $-30° \leq yaw \leq 30°$.

The maximum resolution of the deformation range for translations is 3 $[cm]$, and for rotations is 6°. The total required time for the dataset acquisition was about 3.5 hours. The segmentation and the deformation models are deployed on a laptop with Intel Core i7-7700HQ 2.8 GHz CPU, 16 GB RAM, and NVIDIA GeForce GTX 1050 running Ubuntu 20.04. The deformation estimation routine, including preprocessing segmentation and estimation, runs approximately at 25 Hz, which is close to the hardware bottleneck of the Azure Kinect frame rate. Mobile platform control and manipulator control are deployed on an Intel NUC 10. All software has been developed within the ROS2 framework.

B. Experiments

This Section reports the results of two sets of tests. The first was designed to estimate the performance of our model's observational error. The second was designed to demonstrate that the approach is independent of the human operator and that the human-robot interaction is natural.

1) Observational error: The standard approach in the literature for estimating material deformation in human-robot collaborative transportation relies on tracking the hands' position with motion capture techniques. Among the others, [18] uses a vision-based skeletal tracker for the hands' position and a physics simulator to reconstruct the 3D shape. The simulator design and implementation are relevant for performance achievement in such a family of methodologies, and they change from one use case to another. A fair comparison of such methods with our method is not trivial since implementing the simulator may bias the evaluation.

This set of experiments compares our method with *ad hoc* method we deployed that integrates a motion tracker as the source of input data. Specifically, such a method consists of

1) estimating the deformation status defined in Section II-A using the hand's key points from the official skeleton tracker software provided by the Azure depth camera used in the experiments.
2) The segment connecting the two hand key points is computed, and $\Delta\mathbf{T}$ is computed between the robot grasping point and the middle point of such segment.

This approach provides a fair comparison since, as in [18], the same input data are used and the rest is deterministic. Therefore no other source of error is introduced except for the input data.

We exploited the setup previously used for the dataset acquisition to apply ground truth known deformations to the object. The difference with the dataset setup acquisition is that while the frame holds the ply, the human pretends to hold it without applying further deformations. Specifically,

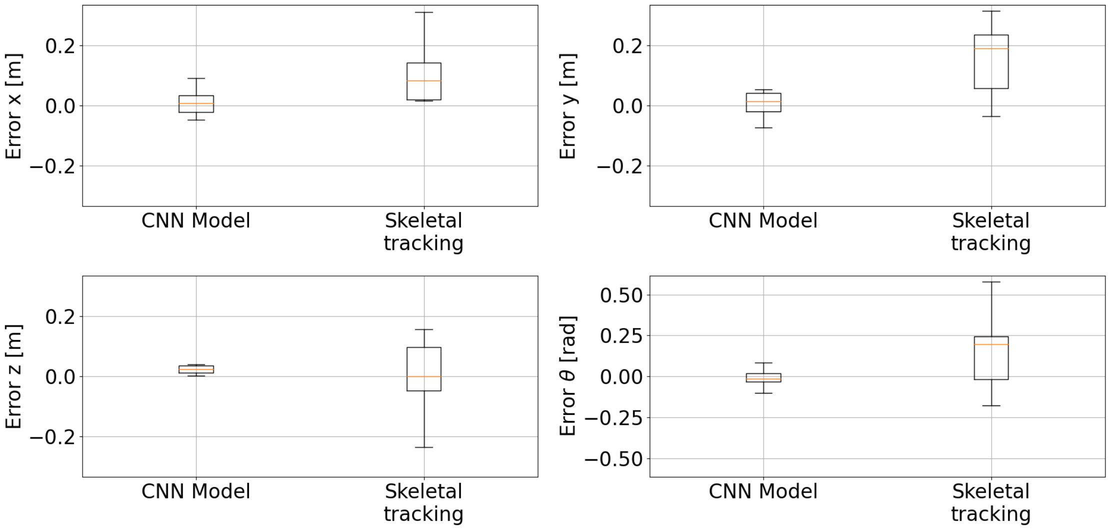

Fig. 6: Comparison of the observational error between our and skeletal tracking methods. **Top left** estimation error on the x-axis. **Top right** estimation error on the y-axis. **Bottom left** estimation error on the z-axis. **Bottom right** estimation error on the z-axis rotation.

the frame pose is acquired via the Apriltags; subsequently, a known deformation is applied, moving the robot to a known relative pose that will be used as ground truth.

Both methods are tested on ten random robot poses, and for each pose, the deformation is estimated ten times; results are shown in Figure 6 as boxplots.

The accuracy of skeletal tracking is generally lower according to the mean estimation error. This depends on the fact that the detected hand key points are placed in the palm's center, not precisely where the ply is grasped. Without a specific calibration, it is impossible to know *a priori* the distance between the key points and the grasping point. It can be noticed that even the repeatability of the measurement, described by the interquartile range, is far worse for the skeletal tracking approach. Indeed the skeletal tracking algorithm produces noisy estimations of the key points, particularly on the edge of the depth image. The error on rotation z is particularly affected. Indeed the ply is pretty tight, and the rotation is computed as the orientation of the segment connecting the hands' key points that, in this case, is not particularly long. Thus a little noise from the hand's key points position can lead to high noise in the estimated rotations. Finally, the skeletal tracker estimation was at a lower rate of approximately 20 Hz.

2) Collaborative transportation: Two human operators, not involved in the human-depth images dataset, were required to perform arbitrary movements in an arbitrary order. The only instructions provided to the operators are: i) the mobile manipulator will follow the human compensating deviations from a predefined deformation status; ii) the mobile manipulator will also try to keep the robotic arm close to a rest configuration. Both the predefined deformation status and the robotic arm's rest configuration were known to the human operators. The video of the experiments is available at the following link: `https://doi.org/10.5281/zenodo.7795589`.

First, the segmentation model proved not to overfit the small number of individuals in the human depth image dataset and was not affected by multiple people in the background and different types of backgrounds. Second, both operators could perform the collaborative transportation naturally without being hindered by the mobile manipulator.

IV. Conclusions

This work presented a data-driven vision-based approach to human-robot collaborative transportation of deformable objects combined with the usage of a mobile manipulator. The proposed approach estimates the deformation status of the comanipulated deformable object via a deformation model based on CNN that takes as input the segmented depth images of it. Depth images are obtained from an RGB-D camera mounted on the robot end effector pointed toward the comanipulated object and subsequently are segmented with an encoder-decoder segmentation model. Both the deformation and the segmentation models are trained on a single dataset whose described acquisition procedure is designed to minimize human intervention. The estimated object deformation is compared with a predefined rest deformation. The delta deformation is first converted into robot twist commands and subsequently converted into joint speed commands considering the system task redundancy. The proposed approach is first compared with the technique for collaborative transportation commonly used in literature based on skeletal tracking. Our approach proved to deliver more reliable measurements in terms of both accuracy and repeatability. Finally, a real case of collaborative transportation was tested with two human operators. The proposed method proved to generalize well, and the operators that had received minimal instruction and no specific training achieved a natural collaboration with the mobile manipulator.

The main drawback of the approach is that a specific dataset for every co-manipulated object, based on shape and mechanical properties, needs to be acquired. Even though the routine is almost fully autonomous, it is still a time-consuming activity, highly limiting its applicability. In future works, the authors plan to use synthetic datasets for training. Furthermore, we will train multiple deformation models concurrently, one for each studied object, using a common CNN backbone to learn more common and general features. Subsequently, applying transfer learning to new comanipulated objects by retraining only the last fully connected layers should require a much lower amount of data.

Acknowledgement

This work is partially supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 101006732, "DrapeBot – A European Project developing collaborative draping of carbon fiber parts."

References

[1] P. Franceschi, N. Pedrocchi, and M. Beschi, "Inverse optimal control for the identification of human objective: a preparatory study for physical human-robot interaction," in 2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, 2022, pp. 1–6.

[2] ——, "Adaptive impedance controller for human-robot arbitration based on cooperative differential game theory," in 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022, pp. 7881–7887.

[3] L. Roveda, S. Haghshenas, M. Caimmi, N. Pedrocchi, and L. M. Tosatti, "Assisting operators in heavy industrial tasks: On the design of an optimized cooperative impedance fuzzy-controller with embedded safety rules," Frontiers in Robotics and AI, vol. 6, 2019.

[4] A. Scibilia, N. Pedrocchi, and L. Fortuna, "Modeling of control delay in human-robot collaboration," in IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society. IEEE, 2022, pp. 1–6.

[5] G. Nicola, E. Villagrossi, and N. Pedrocchi, "Human-robot co-manipulation of soft materials: enable a robot manual guidance using a depth map feedback," in 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2022, pp. 498–504.

[6] J. Sanchez, J.-A. Corrales, B.-C. Bouzgarrou, and Y. Mezouar, "Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey," The International Journal of Robotics Research, vol. 37, no. 7, pp. 688–716, 2018.

[7] D. Andronas, Z. Arkouli, N. Zacharaki, G. Michalos, A. Sardelis, G. Papanikolopoulos, and S. Makris, "On the perception and handling of deformable objects – a robotic cell for white goods industry," Robotics and Computer-Integrated Manufacturing, vol. 77, p. 102358, 2022.

[8] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. H. Adelson, "Cable manipulation with a tactile-reactive gripper," The International Journal of Robotics Research, vol. 40, pp. 1385 – 1401, 2020.

[9] W. Wang and D. J. Balkcom, "Knot grasping, folding, and re-grasping," The International Journal of Robotics Research, vol. 37, pp. 378 – 399, 2018.

[10] A. Verleysen, M. Biondina, and F. Wyffels, "Video dataset of human demonstrations of folding clothing for robotic folding," The International Journal of Robotics Research, vol. 39, pp. 1031 – 1036, 2020.

[11] D. Mcconachie, A. Dobson, M. Ruan, and D. Berenson, "Manipulating deformable objects by interleaving prediction, planning, and control," The International Journal of Robotics Research, vol. 39, pp. 957 – 982, 2020.

[12] A. X. Lee, H. Lu, A. Gupta, S. Levine, and P. Abbeel, "Learning force-based manipulation of deformable objects from multiple demonstrations," in 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 177–184.

[13] Y. Li, Y. Yue, D. Xu, E. Grinspun, and P. K. Allen, "Folding deformable objects using predictive simulation and trajectory optimization," in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 6000–6006.

[14] D. Kruse, R. J. Radke, and J. T. Wen, "Collaborative human-robot manipulation of highly deformable materials," in 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 3782–3787.

[15] B. Jia, Z. Hu, J. Pan, and D. Manocha, "Manipulating highly deformable materials using a visual feedback dictionary," in 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 239–246.

[16] B. Jia, Z. Pan, Z. Hu, J. Pan, and D. Manocha, "Cloth manipulation using random-forest-based imitation learning," IEEE Robotics and Automation Letters, vol. 4, no. 2, pp. 2086–2093, 2019.

[17] D. Sirintuna, A. Giammarino, and A. Ajoudani, "Human-robot collaborative carrying of objects with unknown deformation characteristics," in 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022, pp. 10 681–10 687.

[18] D. Andronas, E. Kampourakis, K. Bakopoulou, C. Gkournelos, P. Angelakis, and S. Makris, "Model-based robot control for human-robot flexible material co-manipulation," in 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 2021, pp. 1–8.

[19] S. Makris, E. Kampourakis, and D. Andronas, "On deformable object handling: Model-based motion planning for human-robot co-manipulation," CIRP Annals, vol. 71, no. 1, pp. 29–32, 2022.

[20] D. D. Schepper, G. Schouterden, K. Kellens, and E. Demeester, "Human-robot mobile co-manipulation of flexible objects by fusing wrench and skeleton tracking data," International Journal of Computer Integrated Manufacturing, vol. 36, no. 1, pp. 30–50, 2023.

[21] P.-C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, "Repeatable folding task by humanoid robot worker using deep learning," IEEE Robotics and Automation Letters, vol. 2, no. 2, pp. 397–403, 2017.

[22] D. Tanaka, S. Arnold, and K. Yamazaki, "Emd net: An encode–manipulate–decode network for cloth manipulation," IEEE Robotics and Automation Letters, vol. 3, no. 3, pp. 1771–1778, 2018.

[23] Y. Tsurumine and T. Matsubara, "Variationally autoencoded dynamic policy programming for robotic cloth manipulation planning based on raw images," in 2022 IEEE/SICE International Symposium on System Integration (SII), 2022, pp. 416–421.

[24] R. Herguedas, G. López-Nicolás, R. Aragüés, and C. Sagüés, "Survey on multi-robot manipulation of deformable objects," in 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 2019, pp. 977–984.

[25] Y. Wang, D. Held, and Z. Erickson, "Visual haptic reasoning: Estimating contact forces by observing deformable object interactions," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 11 426–11 433, 2022.

[26] X. Yu, B. Li, W. He, Y. Feng, L. Cheng, and C. Silvestre, "Adaptive-constrained impedance control for human–robot co-transportation," IEEE Transactions on Cybernetics, vol. 52, pp. 13 237–13 249, 2021.

[27] D. J. Agravante, A. Cherubini, A. Sherikov, P.-B. Wieber, and A. Kheddar, "Human-humanoid collaborative carrying," IEEE Transactions on Robotics, vol. 35, no. 4, pp. 833–846, 2019.

[28] G. Nicola, E. Villagrossi, and N. Pedrocchi, "Co-manipulation of soft-materials estimating deformation from depth images," SSRN Electronic Journal, 2023. [Online]. Available: https://doi.org/10.2139/ssrn.4355722

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269.