Human-robot co-manipulation of soft materials: enable a robot manual guidance using a depth map feedback

Giorgio Nicola, Enrico Villagrossi, Nicola Pedrocchi

***Abstract*— Human-robot co-manipulation of large but lightweight elements made by soft materials, such as fabrics, composites, sheets of paper/cardboard, is a challenging operation that presents several relevant industrial applications. As the primary limit, the force applied on the material must be unidirectional (*i.e.*, the user can only pull the element). Its magnitude needs to be limited to avoid damages to the material itself. This paper proposes using a 3D camera to track the deformation of soft materials for human-robot co-manipulation. Thanks to a Convolutional Neural Network (CNN), the acquired depth image is processed to estimate the element deformation. The output of the CNN is the feedback for the robot controller to track a given set-point of deformation. The set-point tracking will avoid excessive material deformation, enabling a vision-based robot manual guidance.**

I. Introduction

The human-robot co-manipulation of soft materials is becoming a relevant task from the industrial point of view. Several industrial sectors such as aerospace, transport, maritime and renewable energies (*e.g.* wind power and photovoltaic) need lightweight composites materials, usually shaped in large parts challenging to handle and manage using standard automation. The barrier to automation is both technical and economic. First, such parts are delicate, and the processes use irregular raw parts (*e.g.* tissue remnants). Then, the production usually has limited throughput and high product variability. Thus, it is frequent that such production plants employ several operators, which collaborate in handling such lightweight but large parts in a dynamic environment. However, collaboration is often necessary only for a reduced amount of time (*i.e.* transportation of parts), bringing to inefficient production flow.

In such a scenario, the human-robot co-manipulation is full of potential. The human may act as a skilled coordinator and one or multiple robots, or Industrial Mobile Manipulators (IMM)[1], may work as an assistive agent avoiding downtimes and improving the operators' productivity. Specifically, two different human-robot interaction modes are possible. First, the human moves the object in the desired position, and the robot tracks a certain deformation set-point of the material, implementing a robot manual guidance, *e.g.*, suitable for accurate placement of the co-manipulated material. Second, the robot moves along a predefined trajectory while the human tries to follow the trajectory. Suppose the human deviates from the nominal trajectory and the material deformation set-point is not respected. In that case, the robot deviates from the nominal trajectory to keep the material at the desired deformation set-point preventing damages, *e.g.*, suitable for collaborative transport of material from one position to another.

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*

[1]For the nomenclature and acronym refer to RiA R15.08 where Autonomous Mobile Robots (AMR), Industrial Mobile Manipulator (IMM), Autonomous Guided Vehicle (AGV), etcetera, are defined.

A. Related Work

Compared to the manipulation of rigid materials, the manipulation of soft materials introduces new challenges in modeling, perception, grasping, and control [1].

The EU H2020 projects Merging [2] and DrapeBot [3] are pioneering actions coping with these challenges in the industrial scenario. Specifically, the project Merging looks at the manipulation of flexible and fragile objects exploiting multiple industrial robots, designing new Electro-Adhesive (EA) grasping devices, developing new robot AI-based programming and control algorithms supported by the information coming from perception systems fused with the information coming from a digital twin to estimate the deformations of flexible elements [4]. The DrapeBot project focuses on the robotic manipulation of carbon fiber and fiberglass plies during the draping process; in particular, [5] highlights the importance of the human-robot co-manipulation when the dimension of the ply is such as to require more than one robot.

The recent work [6] shows a compelling example of co-manipulation of fabric. The user guides the IMM through gestures recorded by a camera and translated into robot control signals using a skeleton tracking algorithm and force feedback. As a drawback, the user has to execute coded gestures to control the IMM making the cooperation not natural as between two humans during a collaborative task.

Interaction forces cannot be applied through the deformable materials directly to the robot, but only along one direction by pulling the material when it is taut, so force-based sensors are insufficient. Therefore, many authors combined an F/T sensor (*i.e.* mounted between the robot flange and the end-effector) or robot joint torque sensors with a vision system [6], [7]. As a viable solution, [8], [9] and [10] propose the use of visual feedback to detect the material deformations both with 2D and 3D sensors.

Manipulating deformable materials in collaboration with humans or without (often called shape servoing) can be done with model-based and model-free approaches. In model-based approaches, a model, physics-based or black-box, describes the material's mechanical status (*e.g.* deformations,

internal stress, etcetera). In [11] a numerical model is introduced to compute the total elastic energy of the material penalizing deformations from the resting state. The approach is reliable but with a high computational cost. The model in [12] exploits a mass-spring system that uses non-linear springs, dampers, and soft constraints. In [9], after the object shape tracking, thanks to a shape servoing algorithm, the flexible material is manipulated to fit an object template model. The use of a database made by precomputed models of deformable objects (*i.e.* clothes) obtained by simulations is in [13], where each object is modeled as a thin shell to build the synthetic database.

Instead, model-free methods focus on developing handcrafted visual features to be converted directly in robot commands. In [7] a visual feedback controller from the RGB-D image extrapolates the region of the materials with normal vector to much different from the reference. In [8], [14] a set of visual features called Histogram of Oriented Wrinkles (HOW) is developed. In [8] a visual feedback dictionary based on recorded expert users built on top of HOWs is used to compute robot speed commands. Instead, in [14] HOWs are given as input to a Random Forest-based controller trained via imitation learning of an expert user. Finally, in [15] a set of global and local features is created, and an upgraded version of Gaussian Process Regression is used to learn online the controller from feature space to robot velocity commands.

B. Contribution

This paper proposes a data-driven approach to human-robot collaborative manipulation of deformable materials. In detail, a model describing the displacement-deformation relation is learned offline through a neural network. A 3D camera provides the depth map used to train the model. The learned model is used online to determine the displacement from a nominal configuration, and the displacement is fed to a Twist controller. This approach, compared to methods in the literature, has various advantages. First of all, it does not require performing an online physics simulation of the deformable object that is computationally expensive, while computation capability on board of robotic platform is often limited. Second, the implemented controller is very straightforward compared to those described in the literature, such as those based on the deformation Jacobian matrix. Furthermore, it does not require manually developing visual features that might not describe the desired problem fully; instead, the most relevant visual features are autonomously learned during the offline training phase.

The paper is structured as follows: In Section II the problem is formulated and formalized; in Section III the proposed solution is presented; in Section IV the described method is applied to a real setup and experiments and results are presented; in Section V conclusions and future works are detailed.

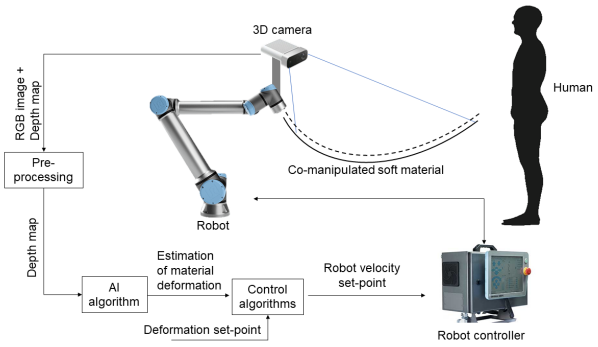


Fig. 1: system concept description

II. Problem Formulation

The problem of human-robot co-manipulation of soft materials is composed of two agents. The first is the human (uncontrollable) that leads the activity, and the second is the robot (controllable) that should adapt to human actions. The objective is to manipulate the desired object while minimizing the deformations from a rest configuration that guarantees not to damage the material. The human brings the co-manipulated material at one side while the robot handles the other side, as shown in Figure 1. Thanks to the controllable agent (*i.e.*, the robot), it is possible to compensate for the inaccuracies introduced by the uncontrollable agent (*i.e.*, the human) and track a set-point of deformation of the co-manipulated material.

Soft materials like textiles can be approximated as membranes [16] characterized by the absence of flexural rigidity and cannot sustain compressive loads. Therefore, deformations can be caused only by displacements or by traction forces. This paper considers the manipulation of fabric, particularly a carbon fiber ply; nevertheless, the method deals with any material approximable with a membrane.

Moreover, manipulating carbon fiber fabric is challenging. Indeed, excessive traction forces can easily damage the fiber structure altering the mechanical property of the material; hence, the material deformations need to be carefully detected. On top of that, carbon fiber is a highly reflective material, and many image analysis techniques struggles. To this purpose, the use of a depth camera, rigidly attached to the robot end-effector, that looks at the top of the co-manipulated element, allows the easy detection of the material shape and deflections due to the forces applied by the human on the material (see Figure 1). The vision sensor can detect material movements along any direction; on the contrary, force-based sensor measure only traction forces, and excessive traction forces can easily damage the material itself.

The 3D camera provides the RGB image and the depth map, and after proper preprocessing, a depth map is obtained composed only by the segmented carbon fiber ply.

The segmented depth map is fed to an ensemble of Convolutional Neural Networks (CNNs) trained to estimate the deformation of the material, in other words, the distance between the robot gripper and the human hands. The

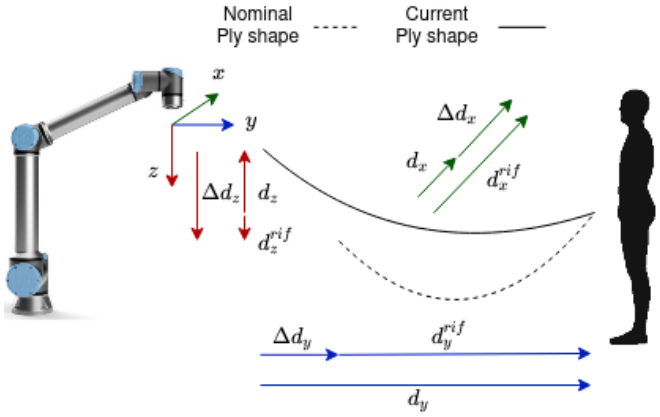


Fig. 2: Problem formulation.

application of Deep Learning techniques allows defining a black-box model describing the relation between visual deformation and mechanical status. This approach is quicker to evaluate than other numerical models, *e.g.*, Finite Element Model (FEM), which are computationally expensive and require additional preprocessing to convert depth images to a compatible input to the model.

Finally, a heuristic transforms the distance estimation into the reference for the robot tool velocity.

As shown in Fig. 2, a nominal shape of the carbon fibre ply is defined as a human-robot reference displacement $(d_x^{rif}, d_y^{rif}, d_z^{rif})$ and the objective is to compute the necessary robot displacement $(\Delta d_x, \Delta d_y, \Delta d_z)$ to reach the nominal shape. We propose a data-driven black-box model composed of an ensemble of CNNs that, given as input a depth image of the carbon fiber ply from the robot point of view, computes the current human-robot displacement (d_x, d_y, d_z).

III. Method

The method described can be divided into four main parts: dataset acquisition, preprocessing, neural network training, and robot control. The paper reports a test case with carbon-fiber plies. Still, the method is general since it does not need any assumption on the material properties but only an RGB-D camera.

A. Dataset Acquisition

The dataset to be acquired consists of multiple depth images of the carbon fiber ply deformed due to the human-robot relative displacement.

Given the poor quality of a long-dataset when humans are grasping the ply for hours in bunches of different configurations, we set up an aluminum frame holding the ply frame, as shown in Fig. 5. Such a frame increases the accuracy and repeatability of the measurements and the robustness of the trained model to human grasping positions variability. The frame, indeed, allows simulating different distances between hands and different inclinations.

The same camera that measures the ply deformation tracks a pair of fiducial markers (Apriltags [17]) placed on the

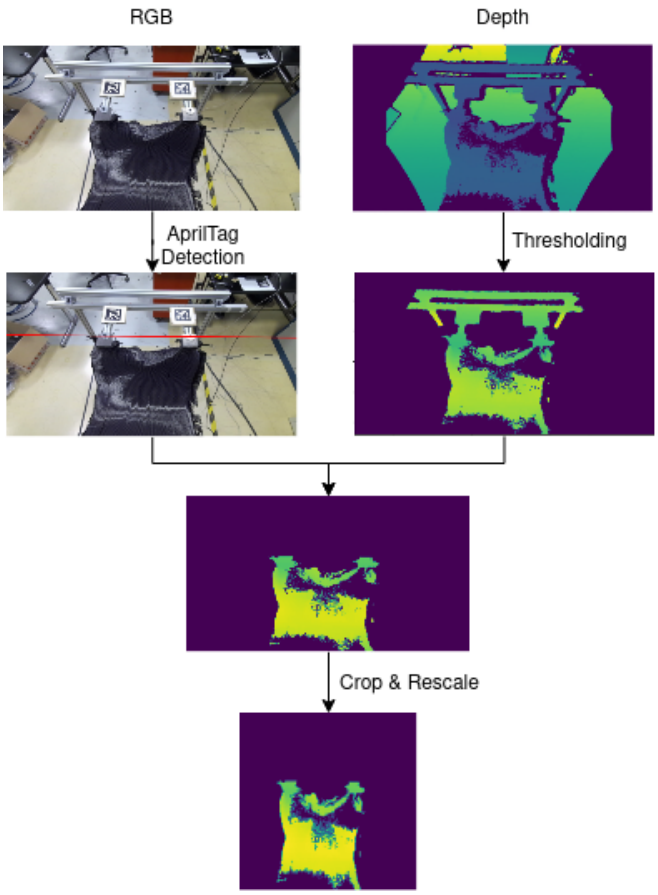


Fig. 3: Example of RGB-D image preprocessing.

frame to detect the frame position relative to the robot, granting high accuracy in the measure of the relative position from the frame. A program based on the motion planning software *MoveIt!* [18] moves the robot relative to the frame in the various studied directions. When the robot is in a new position, we acquire multiple RGB-D images to lower the noisy camera output. The corresponding label is the distance in the three directions for each dataset entry.

B. Preprocessing

The developed preprocessing for the training phase is in Fig. 3. The preprocessing is a two-step algorithm: segmentation and crop plus resize. In detail, in the depth image, the ply is segmented from the background, the aluminum frame used during the dataset acquisition, and the human co-worker during the online phase. First, we set a threshold value of the depth and put each pixel value above the threshold to zero. Concurrently to such elaboration, we convert the RGB image in grayscale and find the AprilTags position. Based on the tag's position, a further threshold line is defined, and we set all pixels above it to zero. During the online phase, instead of using AprilTags, the RGB image is converted in HSV color space, and a range of hue values is defined to segment the carbon fiber ply from the background and the human co-worker.

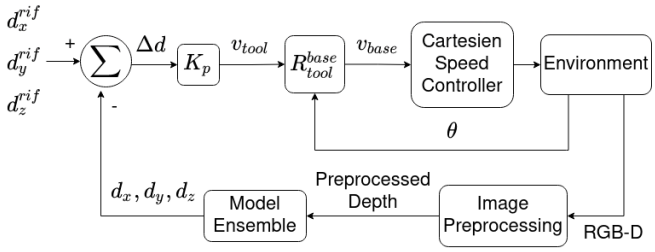


Fig. 4: Control scheme.

Finally, the depth image is cropped and rescaled from the original dimension of 1080×1920 pixels to 128×128. During the online phase, to reduce the noise in input to the CNNs, the last three depth images are averaged, relying on the higher frame rate of the camera has, about 30 Hz, compared to the frequency of the controller, Sect. III-D.

C. Neural Network and Training

The model takes inspiration from VGG16 [19]. The net shares the same general architecture based on blocks of two convolutional layers interspersed by batch normalization and then a maxpool layer. After those blocks, fully-connected layers combined with dropout layers are implemented, and the net output is the cartesian distance along with the three directions. The nets' differences consisted of number blocks, kernel sizes, and strides of both convolutional and maxpool layers.

In detail, we trained three slightly different nets on different subsets of the dataset, given the fact that multiple depth images are taken for each relative robot-human position, The dataset is divided into two separate datasets for each net: training (80%) and test (20%) datasets. Such choice allows verification during the model's training to generalize over unseen relative robot-human positions.

We used Optuna [20] to optimize the hyperparameters of each net, implementing the K-fold cross-validation (K=5). The outputs from the various nets are combined by averaging. As data augmentation, pepper and Gaussian noises and random translations are applied.

D. Robot Control

The robot control scheme is described in Figure 4. After the image preprocessing, the model ensemble (*i.e.* ensemble of CNNs) outputs the estimated human-robot distance, *i.e.*, the ply deformation, and a proportional controller converts the error in ply deformation, Δd, to a tool velocity in the tool frame. To avoid excessive tool velocities, v_{tool} is saturated to a maximum of 5 cm/s in every direction. Finally, the tool velocity is converted from the tool frame to the robot base frame through the rotation matrix R_{tool}^{base} analytically computed from the robot joints angle θ. The control frequency is 7 Hz, higher than the average human reaction time.

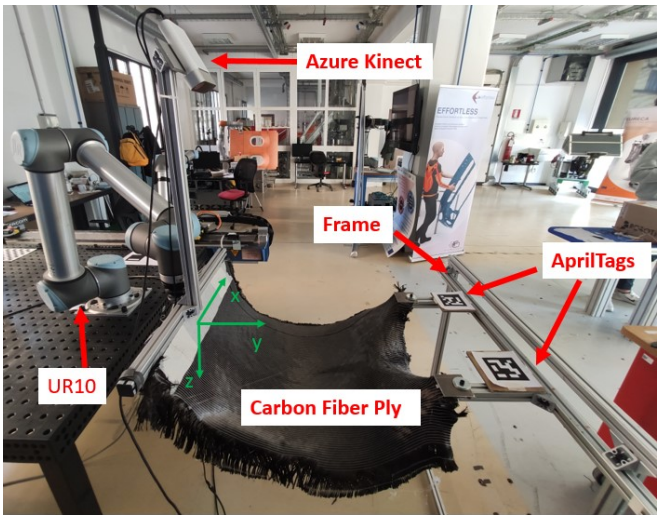


Fig. 5: Setup developed for the dataset acquisition.

IV. Experiments and Results

A. Experimental Setup

The experimental setup is in shown Figure 5 and is composed of an industrial manipulator UR10, an Azure Kinect Camera, and a carbon fiber ply of rectangular shape with dimension 90x65 cm. The range of studied human-robot relative displacement was the following:

- $-0.15 \leq x \leq 0.15$ $[m]$;
- $0.5 \leq y \leq 0.85$ $[m]$;
- $-0.15 \leq z \leq 0.15$ $[m]$.

The displacement range was discretized in all directions with step 0.05 meters during the dataset acquisition. Furthermore, nine possible human hand positions combinations were studied. In conclusion, 455 different human-robot displacements were studied for each human hands position combination, and 2 RGB-D images were taken, with a total dataset dimension of 8190 samples.[2]

The ensemble model used in this application is composed of 3 neural networks, each achieving an average distance error of $(0.0296,\ 0.0275,\ 0.0267)$ $[m]$. After averaging the networks' output, the final average distance error is 0.0215 $[m]$. The proposed model was deployed on a PC desktop with Intel i9-7920X and Nvidia GeForce 3070ti. The total time to preprocess and evaluate each camera frame is 23.65 ms, which is significantly lower than the frame rate of the Azure Kinect (30 Hz), allowing the deployment of the application in real-time. [3]

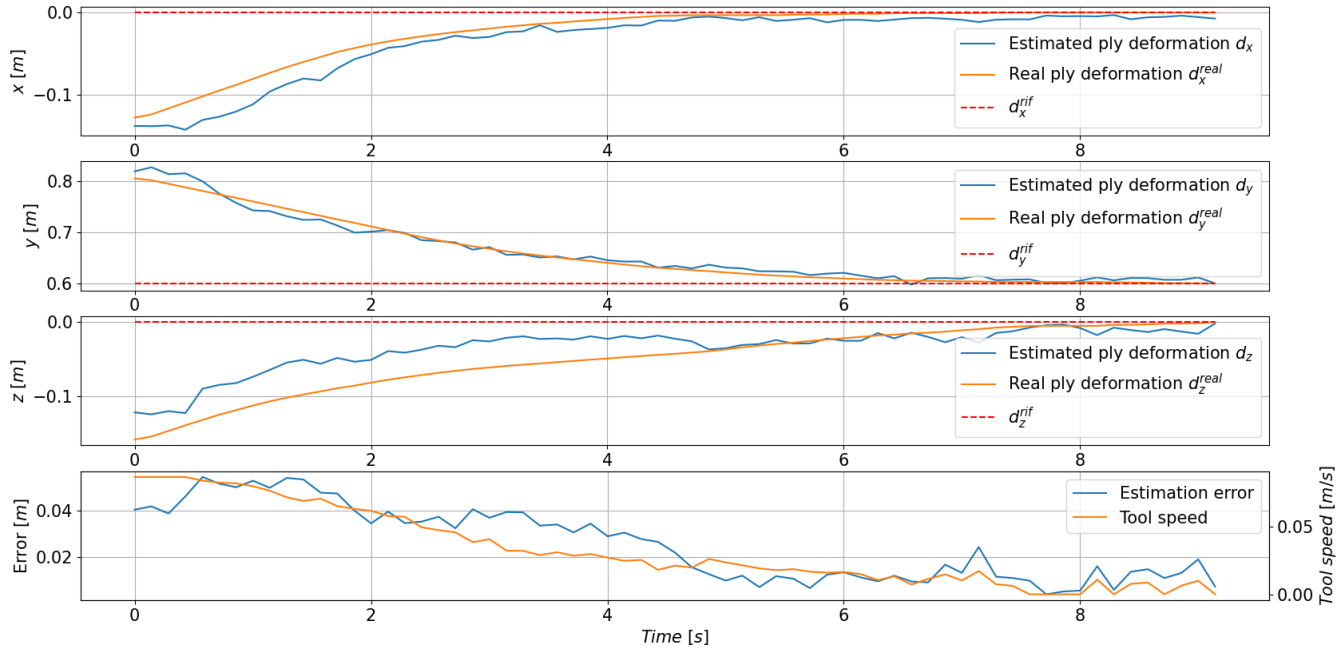
B. Experiments

The Section reports the results of two different tests [4]. First, we analyze the step response to a ply deformation. Then, we analyze a manual guidance operation.

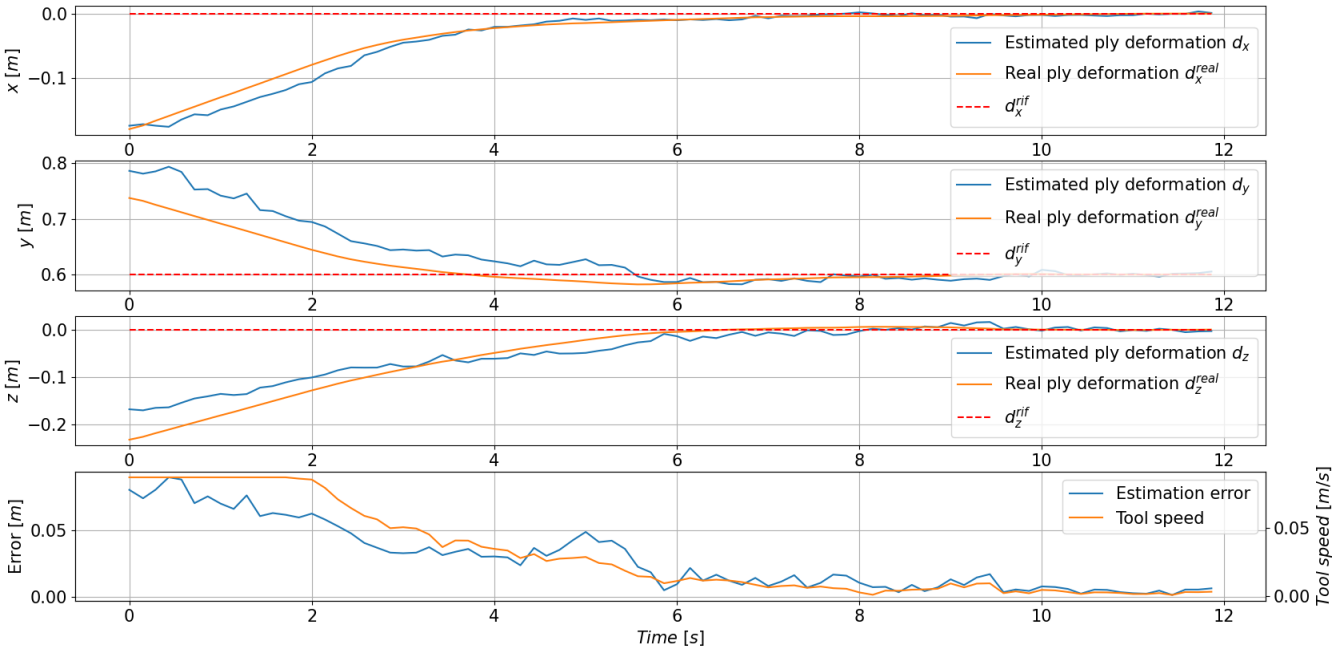
[2]Dataset available at 10.5281/zenodo.6380409

[3]Code for training, testing and deployment available at https://github.com/giorgionicola/SMAHRCO

[4]Video describing the experiments available at https://zenodo.org/record/6379312



(a)



(b)

Fig. 6: Analysis of the step response and estimation error relation to tool speed. for each step response, (a) and (b), it shown the estimated and real ply deformation along the axis x-y-z and the total estimation error compared with the robot tool speed.

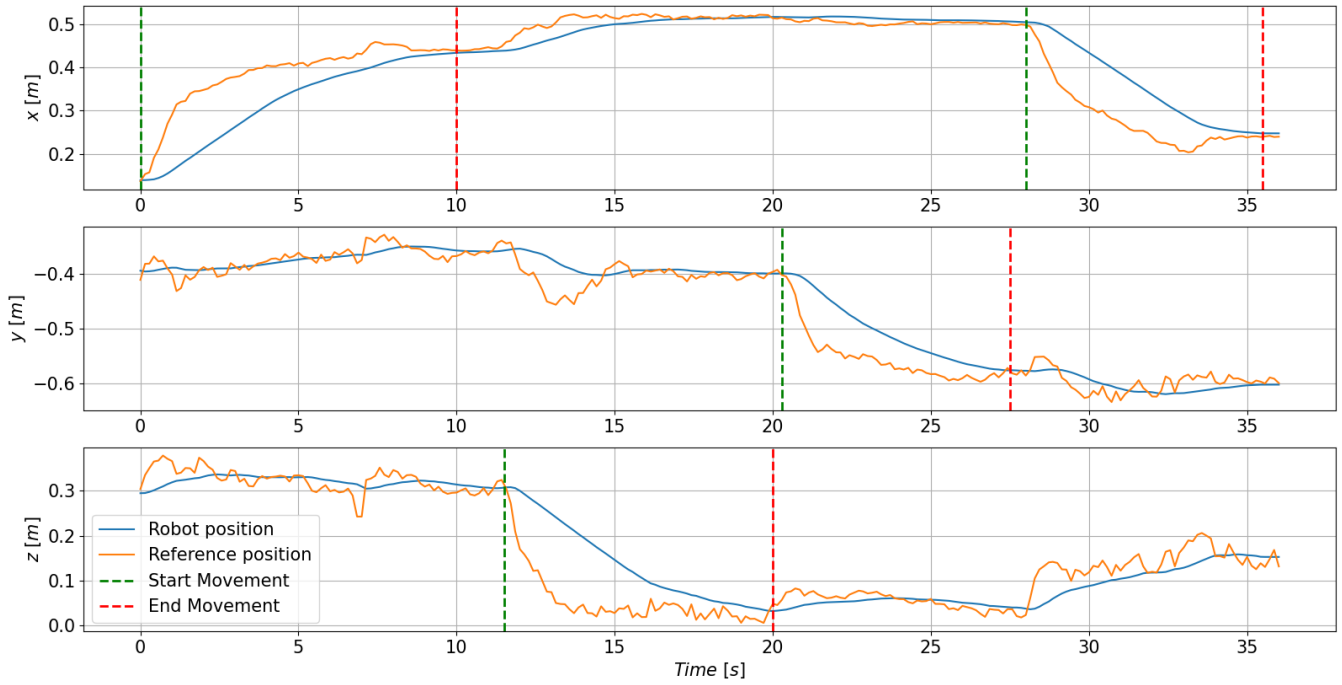


Fig. 7: Analysis of a manual guidance application. The human-robot reference displacement is converted into robot tool reference position (orange line). The robot position is the blue line. Green and red dashed lines highlight the start and end of successive human movement.

1) Step Response Analysis: The ply was attached to the frame used during the dataset acquisition. The robot starting position was displaced of a known value from the ply resting configuration set to $(d_x = 0, d_y = 0.6, d_z = 0)$.

Figures 6a and 6b show the results of two different trials. In both cases, the robot reaches the ply rest configuration successfully.

The bottom graphs in Figures 6a and 6b report the estimation error and its relation with the robot tool speed. The error is maximum at the beginning, decreasing accordingly with the robot tool speed. Indeed, as detailed in Sect. III-D, the vision system (camera and preprocessing) runs approximately at 30 Hz, while the controller receives the averaged last three frames. On the one hand, averaging the depth images reduces noise and, therefore, improves the stability of the estimation of the ply deformation. On the other hand, the average depth image becomes slightly blurred when the robot tool is high, and the estimation accuracy decreases.

Nevertheless, the inaccuracy of the estimation is still consistent with the actual ply deformation, *i.e.*, it never estimates a deformation in the opposite direction of the real one. Furthermore, the developed system can recover from the inaccurate estimation and converge to the desired ply deformation.

Figure 6b shows the estimation robustness when the deformation is beyond the limits of the dataset acquisition campaign. Indeed, the initial deformation in x and z directions are 0.18 and 0.23 meters. Even though the estimation is somehow inaccurate in the z-direction, it is still consistent, and the system can converge to an accurate to the desired ply deformation.

2) Manual Guidance: Finally, in Figure 7 we studied the case of manual guidance. The human was required to perform four movements of arbitrary lengths in the direction $x \rightarrow z \rightarrow y \rightarrow x$ highlighted between the green (start of the human movement) and the red (end of the human movement) dashed lines. The robot could follow human instructions in all cases, and the robot movements were smooth. Even in this scenario, the human could efficiently perform movements that would require the ply to deform beyond the limits in the dataset, confirming the robustness of the approach.

V. Conclusions and Future Works

This paper proposes a Data-driven method for human-robot co-manipulation of flexible materials. The method implements black-box model, based on an ensemble of deep neural networks, that estimates current relative human-robot displacement from depth images. Subsequently, the displacement error to a reference displacement turns into Twist command. The paper also describes the methodology used to acquire the dataset, preprocess it, and train the ensemble model. The proposed method achieved an overall mean average error of 0.0215 $[m]$, and it requires a computation time, including preprocessing, of 23.65 ms, thus allowing to deploy it in real applications. The method was then tested and proved capable of compensating for undesired deformation of the carbon fiber ply both during the analysis of a step deformation response and in a manual guidance application.

Currently, the trained model is limited to movements in the three principal directions x-y-z, while it does not take

into account rotations. Thus, we plan to acquire a dataset including also rotations. The proposed method uses, as input, depth images that are sensitive only to macroscopic deformations; thus, it is not particularly sensitive to traction forces that typically produce much lower intensity deformations. To reduce noise, depth images were averaged with the drawback of increased inaccuracy at higher robot tool speeds. To solve noisy inputs, the samples taken for each robot position during the dataset acquisition will increase, the camera noise during the data augmentation will be simulated faithfully, and regularization during the training will be increased. Traction forces could be helpful to discriminate between movements that produce similar deformations like some translations and rotation. Therefore, we plan to introduce a sensor fusion between the forces and RGB-D images. Finally, the method was tested on a setup with a single industrial manipulator. However, introducing an IMM or a fleet of IMMs in a dynamic environment as a robotic partner would significantly increase the technological fallout of the work.

ACKNOWLEDGEMENT

This project has received funding from 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] 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.

[2] MERGING Consortium, "Manipulation enhancement through robotic guidance and intelligent novel grippers (merging)," 2019. [Online]. Available: http://www.merging-project.eu/

[3] DrapeBot Consortium, "Drapebot – a european project developing collaborative draping of carbon fiber parts," 2021. [Online]. Available: https://www.drapebot.eu/

[4] D. Andronas, G. Kokotinis, and S. Makris, "On modelling and handling of flexible materials: A review on digital twins and planning systems," Procedia CIRP, vol. 97, pp. 447–452, 2021, 8th CIRP Conference of Assembly Technology and Systems.

[5] C. Eitzinger, C. Frommel, S. Ghidoni, and E. Villagrossi, "System concept for human-robot collaborative draping," in SAMPE Europe Conference, 2021, pp. 7542–7549.

[6] D. De Schepper, B. Moyaers, G. Schouterden, K. Kellens, and E. Demeester, "Towards robust human-robot mobile co-manipulation for tasks involving the handling of non-rigid materials using sensor-fused force-torque, and skeleton tracking data," Procedia CIRP, vol. 97, pp. 325–330, 2021, 8th CIRP Conference of Assembly Technology and Systems.

[7] 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.

[8] 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.

[9] M. Aranda, J. Antonio Corrales Ramon, Y. Mezouar, A. Bartoli, and E. Özgür, "Monocular visual shape tracking and servoing for isometrically deforming objects," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 7542–7549.

[10] L. Bodenhagen, A. R. Fugl, A. Jordt, M. Willatzen, K. A. Andersen, M. M. Olsen, R. Koch, H. G. Petersen, and N. Krüger, "An adaptable robot vision system performing manipulation actions with flexible objects," IEEE Transactions on Automation Science and Engineering, vol. 11, no. 3, pp. 749–765, 2014.

[11] D. Kruse, R. J. Radke, and J. T. Wen, "Human-robot collaborative handling of highly deformable materials," in 2017 American Control Conference (ACC), 2017, pp. 1511–1516.

[12] 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.

[13] Y. Li, Y. Wang, Y. Yue, D. Xu, M. Case, S.-F. Chang, E. Grinspun, and P. K. Allen, "Model-driven feedforward prediction for manipulation of deformable objects," IEEE Transactions on Automation Science and Engineering, vol. 15, no. 4, pp. 1621–1638, 2018.

[14] 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.

[15] Z. Hu, P. Sun, and J. Pan, "Three-dimensional deformable object manipulation using fast online gaussian process regression," IEEE Robotics and Automation Letters, vol. 3, no. 2, pp. 979–986, 2018.

[16] V. V. Vasiliev and E. V. Morozov, "Chapter 8 - equations of the applied theory of thin-walled composite structures," pp. 575–590, 2018.

[17] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in Proceedings of the IEEE/RSJ International Conference on Intelligent, Robots and Systems (IROS), October 2016.

[18] D. Coleman, I. Sucan, S. Chitta, and N. Correll, "Reducing the Barrier to Entry of Complex Robotic Software: a MoveIt! Case Study," arXiv e-prints, p. arXiv:1404.3785, Apr. 2014.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2015.

[20] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," 2019.