2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

A Virtual Reality Framework For Fast Dataset Creation Applied to Cloth Manipulation with Automatic Semantic Labelling

Júlia Borràs, Arnau Boix-Granell, Sergi Foix, and Carme Torras

***Abstract*— Teaching complex manipulation skills, such as folding garments, to a bi-manual robot is a very challenging task, which is often tackled through learning from demonstration. The few datasets of garment-folding demonstrations available nowadays to the robotics research community have been either gathered from human demonstrations or generated through simulation. The former have the great difficulty of perceiving both cloth state and human action as well as transferring them to the dynamic control of the robot, while the latter require coding human motion into the simulator in open loop, i.e., without incorporating the visual feedback naturally used by people, resulting in far-from-realistic movements. In this article, we present an accurate dataset of human cloth folding demonstrations. The dataset is collected through our novel virtual reality (VR) framework, based on Unity's 3D platform and the use of an HTC Vive Pro system. The framework is capable of simulating realistic garments while allowing users to interact with them in real time through handheld controllers. By doing so, and thanks to the immersive experience, our framework permits exploiting human visual feedback in the demonstrations while at the same time getting rid of the difficulties of capturing the state of cloth, thus simplifying data acquisition and resulting in more realistic demonstrations. We create and make public a dataset of cloth manipulation sequences, whose cloth states are semantically labeled in an automatic way by using a novel low-dimensional cloth representation that yields a very good separation between different cloth configurations.**

I. Introduction

Research on versatile cloth manipulation by robots is gaining momentum due to the increasing interest in automating daily tasks in assistive contexts. This research is particularly challenging because of the infinite-dimensional space of cloth configurations, in contrast to the 6-dimensional space of rigid object poses, as well as the difficulty of determining how to manipulate a clothing item for fulfilling a given task.

Since full observability is impossible in a real scenario, cloth state needs to be estimated. Whereas the pose of a rigid object can be easily estimated once a portion of its body is identified and located in 3D space, an accurate deformable objects' state is nearly impossible to infer with just partial observability.

The research leading to these results receives funding from the European Research Council (ERC) from the European Union Horizon 2020 Programme under grant agreement no. 741930 (CLOTHILDE: CLOTH manIpulation Learning from DEmonstrations) and project SoftEnable (HORIZON-CL4-2021-DIGITAL-EMERGING-01-101070600). Authors also received funding from project CHLOE-GRAPH (PID2020-118649RB-I00) funded by MCIN/ AEI /10.13039/501100011033 and COHERENT (PCI2020-120718-2) funded by MCIN/ AEI /10.13039/501100011033 and cofunded by the "European Union NextGenerationEU/PRTR".

The authors are with Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028 Barcelona, Spain. {jborras, aboix, sfoix, torras}@iri.upc.edu

(a) (b)

Fig. 1: The manipulation is done on top of a real table that has its virtual version inside of the framework. To make sure that each pair of objects (real and virtual) are located in the same place (in the real and virtual worlds), we used *HTC* trackers for extrinsic calibration. (1a) Real setup showing HTC's tracker (top left), headset (middle) and controllers (bottom). (1b) Virtual setup showing HTC's tracker (top left), controllers (bottom) and simulated garment (center).

Moreover, to decide what grasp to use or which actions to perform to shape cloth in a particular way, several factors need to be taken into account, such as friction, elasticity and thickness of the fabrics or the weight, size and shape of the garment.

The mentioned challenges are probably the reason why there are not as many good datasets of deformable objects (and textiles, in particular) as there are of their rigid counterparts. This fact is slowing down the development of algorithms for the perception and manipulation of cloth-like objects.

One of the main decisions when trying to build a dataset of garment manipulation demonstrations is whether to use real or simulated pieces of fabric. Currently, most of the available datasets are based on RGB-D images coming from real clothing data [1]–[7]. Despite the convenience of having real data, it is very hard to extract the ground truth information from garments and humans along a manipulation sequence. Moreover, data tend to have noise and multiple occlusions, and manual post-processing is always needed in order to have good estimated labeling. Other approaches exploit the use of simulation environments to easily obtain fully observable ground truth data, although they must program the cloth manipulation behaviours with scripted trajectories. Therefore, this type of data lacks human-like demonstrations, losing the crucial variability and manipulation dexterity contributions that would be provided by having the human perception into

the loop. For instance, imagine the movement followed by a human hand previous to the prehension of a deformable object. Having the human in the loop would permit determining whether a grasping point is adequate and what subsequent manipulations should be applied to fulfill the task.

In order to address these challenges, we propose a new approach that combines the use of simulated garments with human manipulation trajectories. Thanks to a virtual reality (VR) framework, humans can interact in real time with simulated pieces of cloth (see Fig. 1a). Using the proposed framework we create a dataset of human cloth folding demonstrations. The dataset can be found in the paper website [1]. In addition, we present a methodology to automatically label cloth states using a novel low-dimensional representation introduced in a previous work by the authors [8]. The labeling of the cloth deformation states during manipulation enables to link high-level planning with low-level features and trajectories of the cloth.

This work is an extension of the paper [9] where preliminary testing of the framework was carried out using a small dataset as a proof of concept. In the current work, we collected an extensive dataset and introduced a new method for labeling cloth states. Note that this automatic labeling method is general and it not only can be used to label the cloth states in the manipulation sequences of our dataset, but it can also be applied to label all the frames collected by means of our framework in future works by other researchers and in other applications.

This article is structured as follows. Section II analyzes the related work in the literature. In Section III we present the different parts of our virtual reality set-up. Section IV introduces the different manipulation sequences used to collect all the data and the way we have organized them. Section V introduces the methodology to automatically label the cloth states, using a novel cloth representation. Finally, Section VI summarizes our contributions and gives prospects for future work.

II. Related Work

In the last years the creation of cloth manipulation datasets has enabled a lot of progress in cloth state estimation and classification. Existing datasets use either real or simulated fabrics in order to provide rich, and as accurate as possible, data reservoirs of cloth types, manipulation actions and garment states distribution.

In the context of garments, several attempts have been made to create various datasets. Some of them classify the garments by type [10]–[14], studying only static properties. Therefore, they are not useful when trying to understand manipulation processes. Others focus on the actions performed by a human when manipulating garments with RGB images [15], and they can learn semantic states as in [16] but it is difficult to link them to low level features of the cloth. RGB-D (or RGB) images are also used in [1]–[7], [17]. These approaches cannot deal with cloth self-occlusions and ground truth information on cloth state is very expensive to obtain, as manual labeling of the images is required.

Other works use reinforcement deep learning directly on simulation [18]–[20], where a specific cloth configuration is the target and the system learns to move a corner of the garment to achieve that configuration. These approaches have shown promising results, but only simple manipulations have been learned, and dynamic motions are still problematic to transfer to real because the simulators cannot accurately estimate the dynamic behaviour of clothes. In addition, such works do not generate datasets to be reused for other applications.

At the time of writing, and despite the variety of proposed approaches, the authors have no knowledge of any other studies that provide both the actions performed by a human while manipulating garments and, at the same time, tracking the full evolution of the piece of fabric from an original state (before manipulation) to an ending state (after manipulation). Our approach aims to fill this void.

Actually, in the robot manipulation community there is a long–standing general agreement that "low-complexity representations for the deformable objects should be the objective" [21] and some attempts to use topological constructs to this end have been made, using writhe matrices, winding numbers, Laplacian coordinates or cell complexes for topology-based representations [22]–[25]. In this work, we will be using a novel low-dimensional representation, the dGLI coordinates [8], which has proven to yield the adequate discrimination granularity between cloth states required in our context.

Similarly, to enable efficient planning of a sequence of actions to take a deformable object from one state to another one, we must also simplify the state-action representation. For instance, classifying and reducing the infinite number of cloth deformation states and possible manipulations that can be applied to them. With this in mind, to build the database we have segmented the executed manipulations in intervals depending on the number of grippers used (one or two), the type of contact (single point P, linear L, or planar Π) following [26] and the part of the garment where the contact is made, following [27] (See Fig. 2). This means we perform tasks using different grasp types, opening the door to study the influence of different grasp types when manipulating cloth, and we reach final states following diverse sequences of manipulations, providing data to learn alternative manipulations to execute one same task.

The developed framework is general and other datasets could be easily recorded, providing the community with a tool to test and compare different ways of organizing and representing manipulation tasks, as depending on the manipulation task performed different representations may be required.

Finally, our study focuses on a simple piece of squared cloth instead of other possible clothing items like T-shirts, pants or jackets because we strongly believe we first need to understand the fundamentals on how to model and classify

[1] http://www.iri.upc.edu/groups/perception/#VR_Framework_Dataset

(a) 1 $\text{PP}^{c} + \Pi_e$ (b) 1 $\text{LL}^{m} + \Pi_e$ (c) 2 $\text{PP}^{c+c} + \Pi_e$ (d) 2 $\text{PP}^{m+m} + \Pi_e$ (e) 2 $\text{PP}^{c+m} + \Pi_e$

Fig. 2: Classification of the different types of garment manipulations studied in this work: (a) One corner double point grasp (PP^{c}) with extrinsic planar contact (Π_e), (b) one middle edge double line grasp (LL^{m}) with (Π_e), (c) two corner double point grasp (PP^{c+c}) with (Π_e), (d) two middle edge double point grasp (PP^{m+m}), and (e) one corner, one middle edge double point grasp (PP^{c+m}) with (Π_e).

deformation states for simple objects and learn to execute actions to navigate through them before generalizing to other shapes with more complex topologies.

III. Framework description

A. HTC Vive Pro

We wanted to design a framework allowing the visualization and storing of the manipulated cloth but also allowing to make realistic recreations of garment manipulations in an interactive experience. With that objective, we used *HTC Vive Pro* in combination with *HTC Tracker* and the *HTC Controllers* as shown in the setup Fig. 1. The tracker eases the connection between the real and the virtual world, making it possible to connect virtual objects with its real counterpart, as long as they have the tracker attached. We are already working on improving the virtual experience fusing it with real world, enabling to provide contact feedback with the environment. In addition, we are building simulated grippers that can be attached to the controllers to provide an interaction with the table similar to the one the robot will have. This will open the door to learn more realistic but robot friendly trajectories that use environmental constraints [28] to achieve the tasks. Such environmental constraints are essential in cloth manipulation, as analyzed in [26].

The controllers send their position and orientation from the real to the virtual world. In addition, they can also send information using their integrated buttons, including one pressure-sensitive trigger and a trackpad. We use the controllers to store grasp state information while recording a manipulation, following the grasping and manipulation framework introduced in [26], [27].

B. Unity

For the development of the framework, we decided to use the *Unity* engine. The *HTC* hardware can easily be connected to *Unity*. For detailed description of the software setup, we refer to our previous work [9].

Unity is a cross-platform game engine that can also be used for simulations, some examples found in [29]–[31]. Here, *Unity* is used to build a framework where the information coming from the HTC Vive Pro system is displayed in a 3D environment. Moreover, the game engine will also work as a data reading and processing tool. For simulating the cloth, we use Obi [32], a particle-based physics plugin for deformable objects such as cloth, fluids, ropes or soft-bodies. Compared to the other physics systems available, we found Obi Cloth allowed much more constraints per cloth. Overall, our *Unity* framework is user-friendly and intuitive, allowing interested research groups to easily reproduce our framework.

IV. Data collection

For a better versatility of the collected data, the conducted experiments have been divided into states. Each experiment starts in one described state and ends into another. For each experiment we keep track, in an XML file, of all the elements involved in the cloth manipulation task at a sampling rate of 10Hz. The file contains the evolution of: the cloth mesh, the coordinates of each particle of the cloth, the position and orientation of each *HTC* controller, the state of its trigger, the position and orientation of the possible grasping points, eight in our example, and whether these grasping points are being grasped by any of the controllers.

In order to keep the dataset to a reasonable size but as rich as possible, we tried to just perform the most representative garment manipulations. We use both single handed and bi-manual interactions, and we used them over different combinations of point, line and plane contact types to realize different grasp types from [26]. Due to the data format, it is easy to filter the manipulations by contact or interaction types with the objective of applying learning algorithms. We show all the manipulation sequences in the form of a graph in Fig. 3. The color code shows the garment manipulations that have been used following notation from [27], detailed in Fig. 2.

As shown in the color legend in the graph, some states can be achieved by performing different types of manipulation. First, nineteen manipulation sequences have been performed, with three repetitions each. These sequences correspond to all the possible combinations of manipulations that start with the top-left state of Fig. 3 and end with one of the states on the right of the image. Then, we executed 66 additional manipulation sequences using the most common grasp types, mainly the PP and double PP, making sure we achieve different shapes for each of the states, depending on which corner is folded first or to which side is folded, left to right and up to down. Some examples of this variability are shown

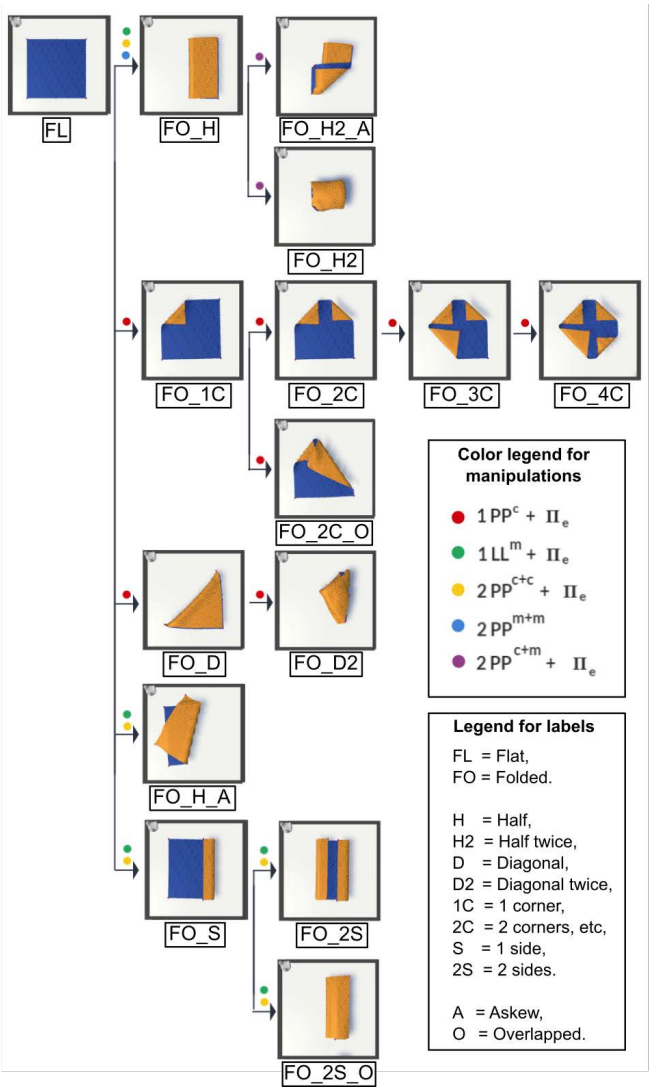

Fig. 3: Graph of state sequences, following the manipulation representation framework in [27]. The coloured dots indicate the different types of grasp types and grasp locations that can be performed to pass from the previous state to the next.

TABLE I: Cloth configurations per label

Label	Possible different samples
FO_1C	
FO_2C	
FO_3C	
FO_S	
FO_H	
FO_H2	
FO_H_A	

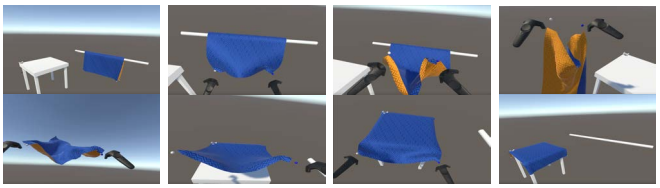

Fig. 4: Manipulation of Tablecloth: From initial hanging position (top-left) to set on the table (bottom-right). The images in-between show frames from the two corner double point grasp manipulation performed on the Tablecloth garment.

in Table I.

We also performed a special case of a manipulation performed with a big *tablecloth* following sizes reported in [33]. The *Tablecloth* garment is hanging from a bar and has to set on the table thanks to a bi-manual manipulation and by taking advantage of the dynamics of the fabrics (see Fig. 4). This task shows how the framework allows to interact with different environmental objects.

V. Semantic Labeling of Cloth States

In order to enable learning of trajectories and folding sequences, it is very useful to have semantic labels that identify different folding or deformation states of the cloth. As each frame in the dataset has information on the grasping state, we can easily group frame intervals of grasped-released cloth states. We could automatically label the released states following the state sequences shown in Fig.3, that have been performed in the human demonstrations. However, due to the nature of human demonstrations, the subject does some extra manipulations to correct mistakes or repeat a grasp, making impossible to automatically label all the dataset just by following the graph of state sequences.

To overcome this issue, we selected only those elements of the dataset that do follow the graph, and used them as a subset from which we can automatically label ground truth semantic labels of cloth state. This subset was used to train a classifier that is then used to label all the rest of the frames in the dataset. To train a good classifier, we need an efficient low-dimensional cloth state representation. To this end, we have used the dGLI *Cloth Coordinates* [8], following a very recent development from the authors. The dGLI Cloth Coordinates can be computed as a closed form formula that depends on the coordinates of 4 pairs of segments of the border of the cloth. In [8] we show how they can be used as a low-dimensional cloth representation that enables a good cluster separation between different cloth configurations.

A. Introduction to the dGLI cloth coordinates

The Gauss Linking Integral (GLI) of two 3D closed and smooth curves is a topological invariant index that measures the linking number between the curves, and when applied twice to the same curve, it measures its writhe or writhing number. Different versions for polygonal curves appeared in the context of DNA protein structures [34] and have been applied in many domains including robotics [23], [24], [35]. Our idea was to directly apply it to the curve formed by the border of the cloth, but in many occasions, when the cloth is on a surface like a table, the curve is planar and the GLI index vanishes. In [8] we introduce the mathematical development of the dGLI that can be applied to planar curves. In short, the dGLI coordinates are computed using a derivative of the GLI of pairs of segments that form the polygonal curve. Given the border of the cloth ϕ, formed by segments $\phi = \{S_i\}$, we select a subset $\phi_{\text{sel}} \subset \phi$ with only 8 of the segments, those close to the corners of the cloth. Then, the dGLI coordinates are

$$C_{\text{dGLI}} = \big(dGLI(S_i, S_j)\big)_{S_i, S_j \in \phi_{\text{sel}}, i>j}. \tag{1}$$

The term $dGLI(S_i, S_j)$ is a derivative of the GLI that depends on a direction of perturbation of the points that form the segments. We chose to perturb the points in the $\vec{e}_3 = (0,0,1)$ direction, as it is orthogonal to the table and the resulting representation will be invariant under rotations and translations on the table. In other words, if a segment is formed by two points $S_i = \vec{AB}$, we denote the perturbed segment as $S_i^* = \vec{AB}^*$ where $B^* = B + \varepsilon\vec{e}_3$. Then

$$dGLI(S_i, S_j) = \frac{GLI(S_i^*, S_j^*) - GLI(S_i, S_j)}{\varepsilon},$$

for a sufficiently small ε, that we have taken as $\varepsilon = 10^{-5}$ in this work. We can now compute the GLI of a pair of segments using the closed form formula introduced in [34]. Let the segments be $S_i = \vec{AB}$ and $S_j = \vec{CD}$, then

$$\begin{aligned} GLI(S_i, S_j) = GLI(\vec{AB}, \vec{CD}) = &\arcsin(\vec{n}_A\vec{n}_D) + \arcsin(\vec{n}_D\vec{n}_B) \\ &+ \arcsin(\vec{n}_B\vec{n}_C) + \arcsin(\vec{n}_C\vec{n}_A) \end{aligned}$$

with

$$\vec{n}_A = \|\vec{AC} \times \vec{AD}\|,\ \vec{n}_B = \|\vec{BD} \times \vec{BC}\|,$$
$$\vec{n}_C = \|\vec{BC} \times \vec{AC}\|, \text{ and } \vec{n}_D = \|\vec{AD} \times \vec{BD}\|.$$

Please, see [8] for extended details on the dGLI derivation.

B. Supervised classification

Given the cloth mesh border points, we can very easily compute the dGLI coordinates of each cloth configuration corresponding to each frame of the simulations. As shown in [8], these coordinates separate very well configurations in classes were the relative position of the selected segments change. Note that each symmetric configuration, as the ones shown in Table I, are separated by the dGLI coordinates in different classes. According to this representation, folding in half leads to configurations at the intersection of several classes, since the relative position of almost all segments are close to a change. That fact, together with the noisy data coming from the simulator due to the human demonstrations, did not lead to good results when applying the same classification method used in [8].

TABLE II: Classifier testing results

	Cloth representation used for training		
Accuracy	**dGLI**	**Norm. border**	**Image**
Test dataset	$98.5\% \pm 0.6\%$	$97.2\% \pm 1.0\%$	$90.7\% \pm 1.3\%$
Rand. P&R dataset	$88.3\% \pm 1.6\%$	$60.3\% \pm 2.1\%$	$16.0\% \pm 1.9\%$

To overcome this fact, we perform a supervised learning classification. To obtain ground truth labeled data, we first eliminated those intervals with very short manipulations that correspond to corrections. Then, we selected those demonstrations that do follow the number of state transitions shown in Fig. 3. Our dataset contains 123 simulations. After removing the short manipulations, a total of 104 simulations follow the graph of state sequences in Fig. 3, containing 315 grasped-released cloth states intervals. As each interval corresponds to a list of frames where the cloth is moving very little after it has been released, many of them correspond to the same cloth mesh. We discarded these frames, except for the initial ones, always corresponding to the flat configuration "FL", for which we do take several equal frames to avoid having very few samples for this class. This results in a total set of 2135 frames of cloth configurations with ground truth labels, using the labels shown under each state in Fig. 3.

We trained 3 classifiers using different representations for the cloth: the dGLI coordinates, the list of normalized points of the border, that is $\mathscr{B} = \{p_i - p_1, i = 1, \ldots n\}$ and an image representation of the border, as the ones shown in Table I. Each sample is labeled with the ground truth label and we then train a random forest classifier method 30 times, with random splits of 67% of the data for training and 33% for testing. We used different classification methods and got similar results. The mean accuracy from the 30 repetitions for each representation used are reported in Table II. As shown in the table, the accuracy for the trained dataset is similar for all representations, because supervised learning is very successful at learning mappings. To test generalization, we executed a new set of 81 simulations with randomized positions and rotations of the cloth. We processed the randomized set in a similar way to obtain the ground truth labels, resulting in a new set of 720 labeled samples. The classifier trained with the dGLI representation, without any retraining, is able to achieve more than 88% accuracy, the normalized border performs 20% worse, and the image training does not work. Results are reported in the second row of Table II.

We show the confusion matrices of the best obtained classifiers on the test set and on the random Pos&Ori dataset in Fig. 5. Results show that the dGLI coordinates are more efficient to classify cloth states, and generalize better. However, we think they could perform even better. Due to the nature of the dGLI coordinates, all the states like folded in half, folded twice in half, etc, are close to a singularity of the dGLI, that is, close to a change of sign of one of the coordinates. As a result, we need a lot of samples to obtain a

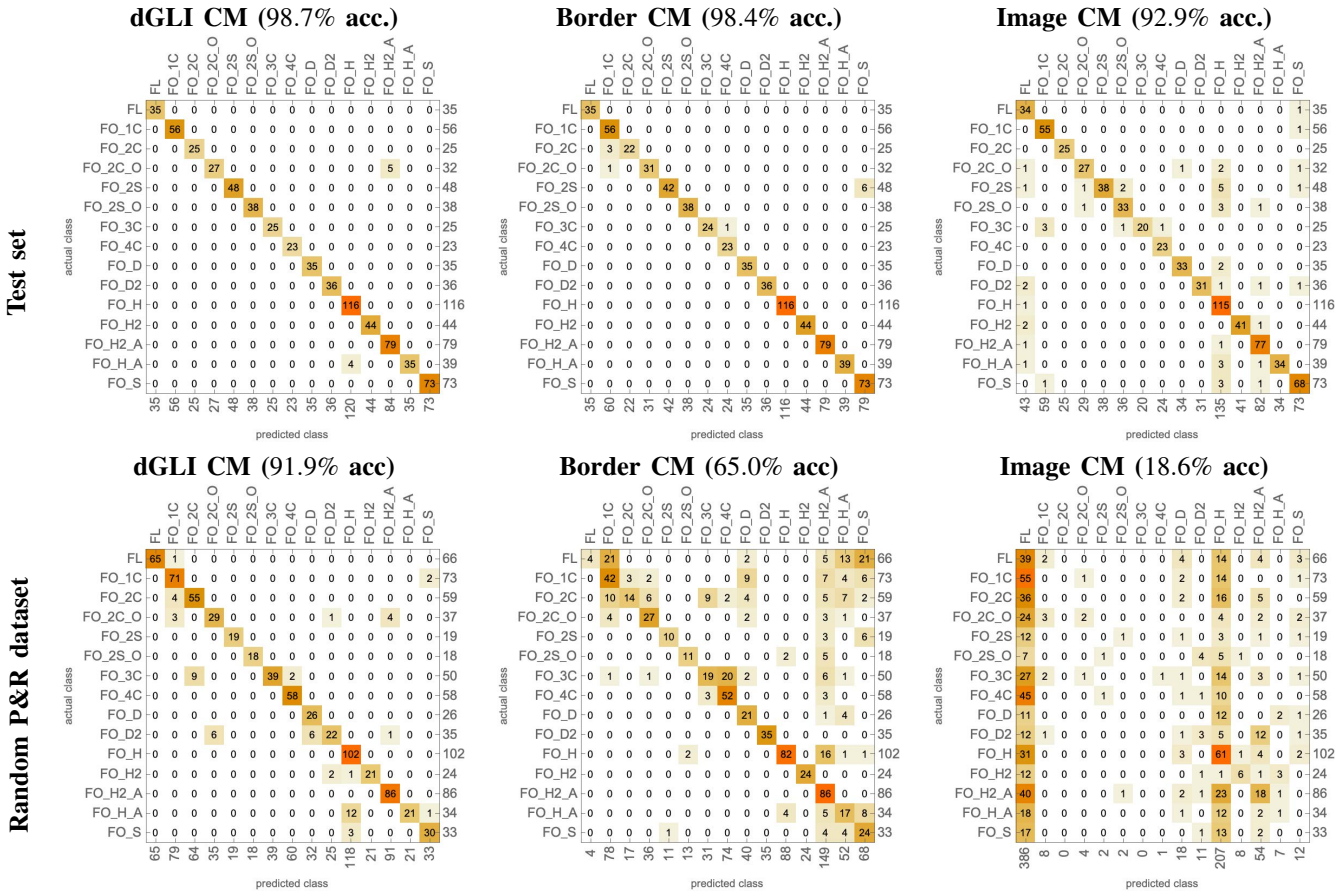

Fig. 5: Confusion matrices (CM) of the three classifiers trained with the three tested representations: the dGLI cloth coordinates, the list of normalized border points and the image representation. In the first row, we show results on the 20% split test of our database. In the second row, we show results for the dataset gathered with random initial position and orientation of the cloth. Our proposed representation clearly outperforms the others on the second dataset.

good variety of relative positions of those states to ensure we have seen representative states all around the singularity. In future work, we will study additional parameters we can add to the representation to solve this issue. Preliminary results on this direction show that if this is solved, a lot less data is needed to achieve similar accuracy. The dGLI coordinates is a novel low-dimensional representation that is based on low-level features of the cloth and can generalize better than other representations like those based on images.

Note that thanks to the manipulation representation used, following [27], we didn't need to manually annotate any data to obtain ground truth of a good percentage of the total dataset. The classifier with best accuracy was then used to label the full database. The labeled dataset is available in the paper website, with other additional material.

VI. Conclusions

In this work, we presented a *Unity* virtual reality framework to perform cloth manipulation experiments. The approach differs from others in that we not only perform a full-mesh tracking but we also keep track of the position, rotation and interactions of other key features of the manipulation (like grippers or grasping points). Moreover, the implementation of the virtual reality allows the creation of an immersive experience that gets rid of the gap between the human and robot perception-action loop.

We use the developed framework to create a rectangular cloth manipulation dataset which is then labeled into high-level semantic states to allow a more versatile study. The use of a novel low-dimensional representation of the cloth allows to train a classifier that outperforms others trained with other usual representations like silhouettes or images.

As future work, we are planning on fusing real contact feedback of the environment with the virtual world to obtain even more realistic manipulations, and on further enhancing the dGLI coordinates to improve the automatic labeling of the high-level state of deformation of the cloth.

The presented fully labeled dataset aims to help the cloth manipulation robotics research community by providing more realistic human-like cloth manipulation data, which can be used in different learning approaches.

References

[1] K. Yamazaki and M. Inaba, "Clothing classification using image features derived from clothing fabrics, wrinkles and cloth overlaps,"

in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2710–2717, IEEE, 2013.

[2] I. Mariolis and S. Malassiotis, "Matching folded garments to unfolded templates using robust shape analysis techniques," in *International Conference on Computer Analysis of Images and Patterns*, pp. 193–200, Springer, 2013.

[3] A. Doumanoglou, A. Kargakos, T.-K. Kim, and S. Malassiotis, "Autonomous active recognition and unfolding of clothes using random decision forests and probabilistic planning," in *2014 IEEE International Conference on Robotics and Automation*, pp. 987–993, IEEE, 2014.

[4] B. Willimon, I. Walker, and S. Birchfield, "A new approach to clothing classification using mid-level layers," in *2013 IEEE International Conference on Robotics and Automation*, pp. 4271–4278, IEEE, 2013.

[5] G. Tzelepis, E. E. Aksoy, J. Borràs, and G. Alenyà, "Semantic state estimation in cloth manipulation tasks," *arXiv preprint arXiv:2203.11647*, 2022.

[6] A. Ramisa, G. Alenya, F. Moreno-Noguer, and C. Torras, "Learning rgb-d descriptors of garment parts for informed robot grasping," *Engineering Applications of Artificial Intelligence*, vol. 35, pp. 246–258, 2014.

[7] E. Corona, G. Alenya, A. Gabas, and C. Torras, "Active garment recognition and target grasping point detection using deep learning," *Pattern Recognition*, vol. 74, pp. 629–641, 2018.

[8] F. Coltraro, J. Fontana, J. Amorós, M. Alberich-Carramiñana, J. Borràs, and C. Torras, "The dGLI cloth coordinates: A topological representation for semantic classification of cloth states," *arXiv preprint arXiv:2209.09191*, 2022.

[9] A. Boix-Granell, S. Foix, and C. Torras, "Garment manipulation dataset for robot learning by demonstration through a virtual reality framework," in *International Conference of the Catalan Association for Artificial Intelligence*, 2022.

[10] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[11] T. Ziegler, J. Butepage, M. C. Welle, A. Varava, T. Novkovic, and D. Kragic, "Fashion landmark detection and category classification for robotics," in *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pp. 81–88, IEEE, 2020.

[12] H. Zhu, Y. Cao, H. Jin, W. Chen, D. Du, Z. Wang, S. Cui, and X. Han, "Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images," in *European Conference on Computer Vision*, pp. 512–530, Springer, 2020.

[13] L. Sun, G. Aragon-Camarasa, S. Rogers, R. Stolkin, and J. P. Siebert, "Single-shot clothing category recognition in free-configurations with application to autonomous clothes sorting," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 6699–6706, IEEE, 2017.

[14] L. Sun, S. Rogers, G. Aragon-Camarasa, and J. P. Siebert, "Recognising the clothing categories from free-configuration using gaussian-process-based interactive perception," in *2016 IEEE International Conference on Robotics and Automation*, pp. 2464–2470, IEEE, 2016.

[15] 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, no. 9, pp. 1031–1036, 2020.

[16] A. Verleysen, M. Biondina, and F. Wyffels, "Learning self-supervised task progression metrics: a case of cloth folding," *Applied Intelligence*, pp. 1–19, 2022.

[21] C. Smith, Y. Karayiannidis, L. Nalpantidis, X. Gratal, P. Qi, D. V. Dimarogonas, and D. Kragic, "Dual arm manipulation—a survey," *Robots and Autonomous Systems*, vol. 60, no. 10, pp. 1340–1353, 2012.

[17] G. Aragon-Camarasa, S. B. Oehler, Y. Liu, S. Li, P. Cockshott, and J. P. Siebert, "Glasgow's stereo image database of garments," *arXiv preprint arXiv:1311.7295*, 2013.

[18] R. Jangir, G. Alenya, and C. Torras, "Dynamic cloth manipulation with deep reinforcement learning," in *International Conference on Robotics and Automation*, pp. 4630–4636, 2020.

[19] D. Seita, A. Ganapathi, R. Hoque, M. Hwang, E. Cen, A. K. Tanwani, A. Balakrishna, B. Thananjeyan, J. Ichnowski, N. Jamali, *et al.*, "Deep imitation learning of sequential fabric smoothing policies," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 9651–9658, 2020.

[20] J. Hietala, D. Blanco-Mulero, G. Alcan, and V. Kyrki, "Learning visual feedback control for dynamic cloth folding," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022.

[22] V. Ivan, D. Zarubin, M. Toussaint, T. Komura, and S. Vijayakumar, "Topology-based representations for motion planning and generalization in dynamic environments with interactions," *The International Journal of Robotics Research*, vol. 32, no. 9-10, pp. 1151–1163, 2013.

[23] W. Yuan, K. Hang, H. Song, D. Kragic, M. Y. Wang, and J. A. Stork, "Reinforcement learning in topology-based representation for human body movement with whole arm manipulation," in *International Conference on Robotics and Automation*, pp. 2153–2160, IEEE, 2019.

[24] F. T. Pokorny, J. A. Stork, and D. Kragic, "Grasping objects with holes: A topological approach," in *IEEE International Conference on Robotics and Automation*, pp. 1100–1107, IEEE, 2013.

[25] F. Strazzeri and C. Torras, "Topological representation of cloth state for robot manipulation," *Autonomous Robots*, vol. 45, no. 5, pp. 737–754, 2021.

[26] J. Borràs, G. Alenyà, and C. Torras, "A grasping-centered analysis for cloth manipulation," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 924–936, 2020.

[27] I. Garcia-Camacho, J. Borràs, and G. Alenyà, "Knowledge representation to enable high-level planning in cloth manipulation tasks," *ICAPS 2022 Workshop on Knowledge Engineering for Planning and Scheduling*, 2022.

[28] C. Eppner, R. Deimel, J. Álvarez-Ruiz, M. Maertens, and O. Brock, "Exploitation of environmental constraints in human and robotic grasping," *International Journal of Robotic Research*, vol. 34, no. 7, pp. 1021–1038, 2015.

[29] T. Ward, A. Bolt, N. Hemmings, S. Carter, M. Sanchez, R. Barreira, S. Noury, K. Anderson, J. Lemmon, J. Coe, *et al.*, "Using unity to help solve intelligence," *arXiv preprint arXiv:2011.09294*, 2020.

[30] A. Juliani, V. Berges, E. Vckay, Y. Gao, H. Henry, M. Mattar, and D. Lange, "Unity: A general platform for intelligent agents. arxiv 2018," *arXiv preprint arXiv:1809.02627*.

[31] M. Honari, "Unity-technologies ml-agents." `https://github.com/Unity-Technologies/ml-agents`, 2013.

[32] V. M. Studio, "[online] obi documentation." `http://obi.virtualmethodstudio.com/`, 2022.

[33] I. Garcia-Camacho, J. Borràs, B. Calli, A. Norton, and G. Alenyà, "Household cloth object set: Fostering benchmarking in deformable object manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 5866–5873, 2022.

[34] K. Klenin and J. Langowski, "Computation of writhe in modeling of supercoiled DNA," *Biopolymers*, vol. 54, no. 5, pp. 307–317, 2000.

[35] S. L. Ho, *Topology-based character motion synthesis*. PhD thesis, University of Edinburgh, 2011.