Video Dataset of Human Demonstrations of Folding Clothing For Robotic Folding

Andreas Verleysen[1], Matthijs Biondina[2] and Francis wyffels[1]

Abstract

General-purpose cloth folding robots do not yet exist due to the deformable nature of textile, making it hard to engineer manipulation pipelines or learn this task. In order to accelerate research for the learning of the robotic folding task, we introduce a video dataset of human folding demonstrations. In total, we provide 8.5 hours of demonstrations from multiple perspectives leading to 1000 folding samples of different types of textiles. The demonstrations are recorded on multiple public places, in different conditions with a diverse set of people. Our dataset consists of anonymized RGB images, depth frames, skeleton keypoint trajectories, and object labels. In this paper, we describe our recording setup, the data format and utility scripts which can be accessed at `https://adverley.github.io/folding-demonstrations`.

Keywords

Deformable objects, robotic manipulation, clothing, learning from demonstration, crowdsourcing

Introduction

Deep reinforcement learning is being applied to many robotic manipulation problems such as grasping everyday objects (Levine et al. 2018), peg-hole insertion (Finn et al. 2016), and bin-picking tasks (Mahler et al. 2019). However, applications concerning the manipulation of deformable objects are scarce due to their highly complex behavior caused by deformations (Foresti and Pellegrino 2004). Clothing, in particular, is a problem relevant to household robotics and industry. While existing work engineers highly complex pipelines (Doumanoglou et al. 2016; Maitin-Shepard et al. 2010) with predefined end-effector trajectories (Miller et al. 2012), some authors recently tried tackling the problem by learning the required manipulation skills (Matas et al. 2018; Seita et al. 2018). These methods require many demonstrations and are trained in simulation, which can lead to transferability issues. Additionally, defining a reward function for the robotic folding problem is hard because of the high dimensionality of the state of textile objects.

Prior work has shown that it is possible to bootstrap learning by starting from task demonstrations (Večerík et al. 2017) or learn the reward function from expert samples (Abbeel and Ng 2004). Learning skills from human demonstrations also provides a communication medium from human to robot, enabling non-experts to train robots or learn from video websites. Datasets containing task executions are useful in this regard as they help research in learning from human demonstrations. For example, the MIME dataset (Sharma et al. 2018) consists of video demonstrations of 20 different tasks. However, only one task deals with deformable objects, the video stream is single-view RGB-D data, and the task consists of wiping with a cloth which is considerably less complicated compared to folding clothing. Datasets that deal exclusively with clothing, for example, DeepFashion (Liu et al. 2016) mainly focuses on cloth recognition tasks making them hard to use for learning the robotic folding task.

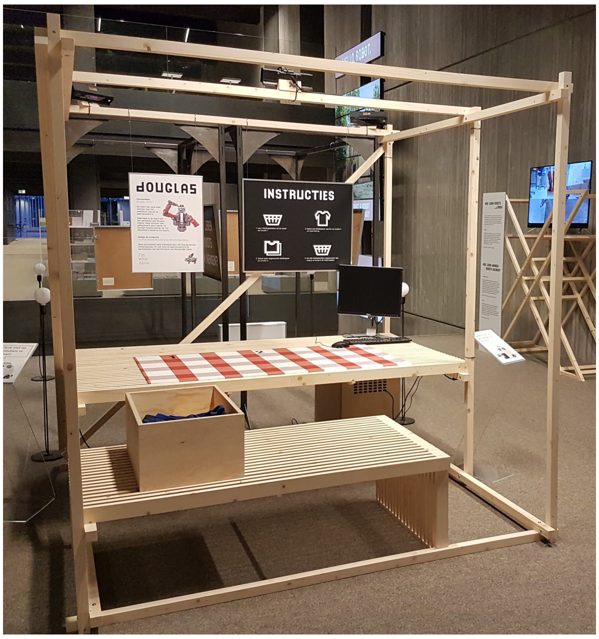


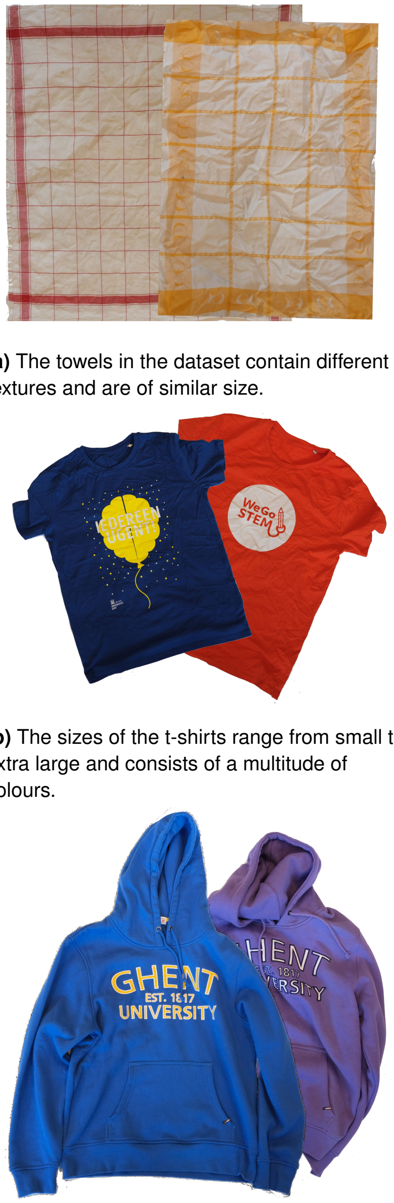
Figure 1. Picture of our folding table setup to crowdsource video demonstrations in a public library.

To fill the gap in learning the robotic folding task from human demonstrations, we provide a dataset that aims

[1] IDLab-AIRO, Electronics and Information Systems Department, Ghent University – imec, Belgium
[2] Radboud University, Netherlands.

Corresponding author:
Andreas Verleysen, IDLab-AIRO, Ghent University – imec, iGent Tower - Technologiepark-Zwijnaarde 126, B-9052 Ghent, Belgium
Email: andreas.verleysen@ugent.be

to accelerate research by providing crowdsourced video demonstrations of people folding clothing. Kinect cameras mounted on a folding table recorded the task executions from multiple perspectives. This resulted in 1000 folding demonstrations in the wild. We segmented and labelled the data such that a subtask of each demonstration can be queried. We provide anonymized RGB images and depth values from three fixed perspectives, together with the skeleton keypoint trajectories. We are confident our dataset of video demonstrations of people folding clothing will accelerate research for learning of the robotic folding task and can be used for action recognition tasks on multiple temporal levels, for example detecting different folding methods and detecting the steps associated with folding textile.



(a) The towels in the dataset contain different textures and are of similar size.

(b) The sizes of the t-shirts range from small to extra large and consists of a multitude of colours.

(c) The hoodies are arguably the hardest piece of textile to fold in the set. There are two hoodies with a different colour.

Figure 2. The set of textiles that has to be folded by the participants consists of hoodies, t-shirts and towels with a variety of sizes, colours and textures.

Crowdsourcing Folding Demonstrations

We gathered a heterogeneous dataset of folding demonstrations using a community-based participatory approach (English et al. 2018). We involve citizens by requesting them to demonstrate their method to fold clothing on a folding table with cameras. Using posters, an instructional video and warning symbols around the folding table setup, we made it explicit that participants will be recorded on video on video for the purpose of research in the domain of robotics and AI. We collected no demographics or other kind of personal information. This setup allows us to capture different folding strategies and manipulation varieties within a folding method. The participants consist of a combination of students and visitors of a public library in the third largest city in Belgium. This avoids selection bias in the dataset. Furthermore, we place our setup within a small exhibition on research in robotics to inform the public about learning strategies for robots and give an answer to an innate fear in society that self-learning robots could lead to a loss in jobs (Fleming 2019).

To capture video task demonstrations, a special-purpose folding table was designed and constructed which can be seen in Figure 1. The table is a beam-like, wooden skeleton structure consisting of a tabletop, a bench, camera mounting points, a basket, and a locker. The participant is required to fold the clothing on the working surface. The tabletop is detachable in order to apply different tablecloths as a means to introduce additional variety in the dataset. As we require the demonstrator to sit while performing the task, we place a large bench attached to the wooden frame. The bench also obstructs observers to prevent occlusion and distraction during task execution. There are three Kinect v2 cameras mounted on top of the table. They capture the perspective from the task executor and two top corner video streams to deal with occlusion. They are placed approximately 160 cm and 183 cm from the center of the folding table in order to capture the complete folding sequence demonstration. The Kinect cameras provide RGB and depth information at a resolution of respectively 1920x1080 and 512x424 pixels. The wooden basket is attached to the bench and serves as a proxy for a laundry basket. Finally, a locker safeguards the workstation embedded in the table. We use the libfreenect2 driver (Xiang et al. 2016) to capture the frames and process the six video streams, RGB and depth information, online using an AMD Ryzen 1700X CPU. Because of the high bandwidth requirements of the Kinect cameras, we limit the frame rate to 10 FPS.

To structure the participants' task demonstrations, we provided a four-step instruction list: (1) place randomly selected clothing out of the basket on the left side of the table, (2) fold one textile at the time in the middle of the table, (3) collect it at the right of the table and (4) put all textile back in the basket. We made an instructional video and put up a poster containing these instructions to avoid high variance in task execution.

Because the folding table is stationed in a public space with mostly no human supervision, we leave it recording throughout the project. To avoid running out of storage and filtering frames without human activity, we run an activity detection heuristic based on changes in the pixel values of

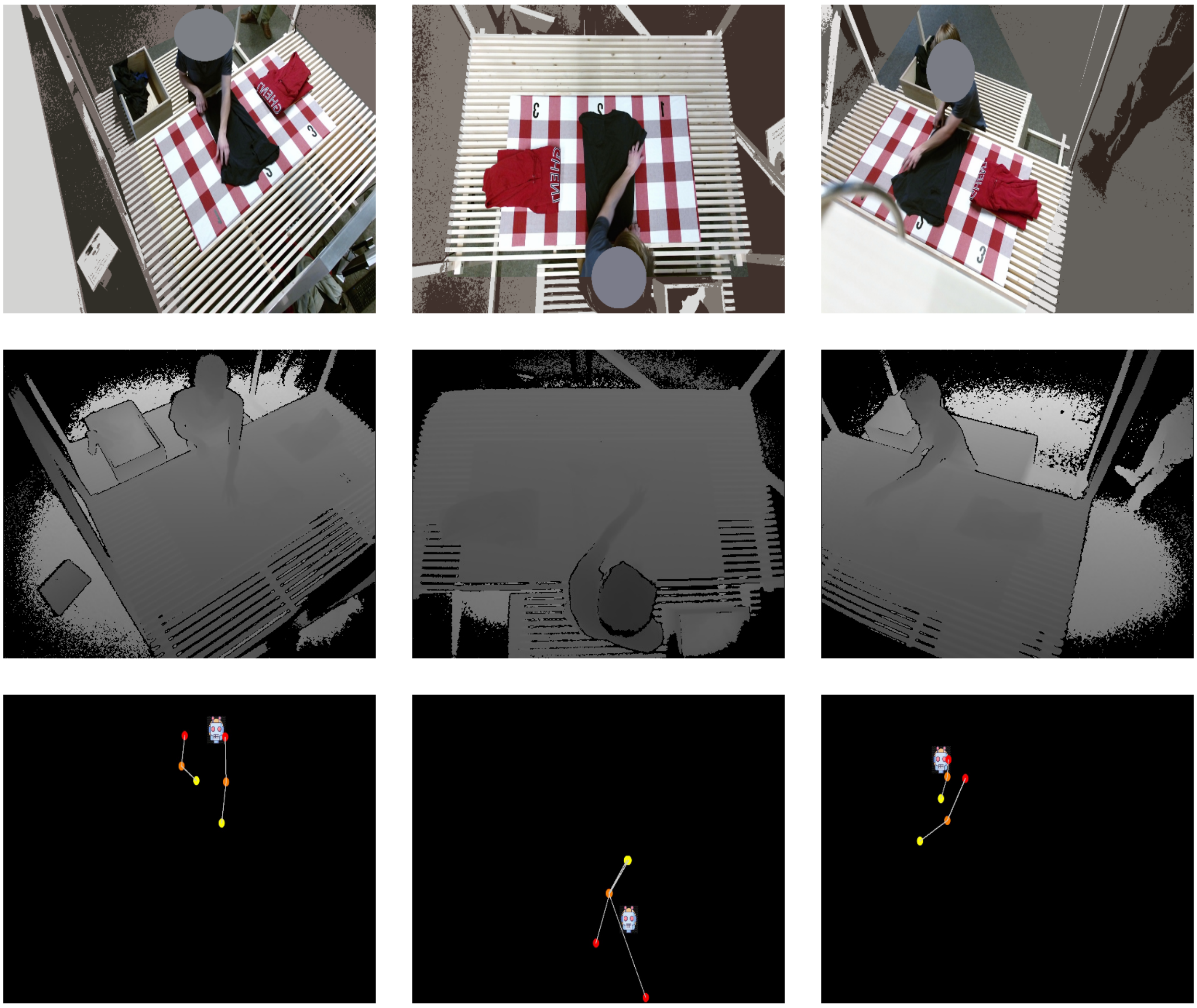


Figure 3. Example output from our dataset. We provide RGB images, depth registrations and skeleton keypoint trajectories of 6.5 hours of human demonstrations of folding clothing. The RGB images are anonymized without compromising image fidelity or disturbing the folding demonstration. The videos are segmented such that one sample represents the folding of one piece of textile.

the video stream. To guarantee ample observations, we also actively visited the setup to attract and inform visitors about the project. We noticed our presence had a positive effect on the number of visitors willing to participate in the data crowdsourcing project.

The included types of textiles in the basket are towels, t-shirts, and hoodies. Examples are shown in Figure 2. We excluded trousers as they are hard to flatten from a sitting position. Socks were also excluded from the set because folding socks require high-dexterous, fine manipulation actions that would not be visible from the mounting position of the cameras.

After capturing the example demonstrations, we cleaned all false positive recordings from the database, sliced the recordings into single-piece folding demonstrations and manually labelled subtasks. The subtasks consist of grasping isolated clothing, unfolding, flattening, folding and stacking it on top of each other. We defined exact definitions of these subtasks in Table 1 in order to consistently label the video fragments. As data quality plays an important role for learning algorithms, we annotated the data ourselves to ensure there is consistency in the labelling between samples. These subtask labels can, for example, be used in reinforcement learning for reward engineering or hierarchical learning and for the training of action recognition systems. Skeleton keypoints were extracted from all frames in post-processing using AlphaPose (Fang et al. 2017).

Folding Demonstrations Dataset

The observations in the datasets are captured over the course of two months at two different public locations. The set contains 1000 folding demonstrations of three different types of textiles. This amounts up to 8.5 hours of folding recorded in 304820 frames. We registered four different types of folding methods. We segmented each video into chunks of single folding demonstrations and provide RGB frames, depth information, annotations, pose trajectories and timestamps of the different steps in the folding task. The content and how to access and use our dataset is described in the remainder of this section.

Subtask	Definition
Isolated grasping	The subject selects grasping points to remove a piece of textile from a heap of multiple textile pieces and isolates the selected textile.
Unfolding	The subject selects grasping points and executes manipulation trajectories in order to remove a fold.
Flattening	The subject executes manipulations in order to remove wrinkles from a piece of textile which can be in any state.
Folding	The subject selects grasping points and executes manipulation trajectories in order to bring the textile into a folded configuration.
Stacking	The subject grasps the textile and moves it outside the folding area, possibly stacking it on top of a pile of folded textiles.

Table 1. Definitions used to label the subtasks in the folding task

Folder structure

The dataset is segmented into folding demonstrations of a single piece of textile. This structure is visible in the folder hierarchy in Figure 4. For each example demonstration, we find annotations and timestamps indicating the subset of the task. Because we have three cameras mounted on fixed positions on the folding table, we put the colour, depth and pose registrations in the folders *left*, *middle* and *right* which represent the viewpoint in front of the table.

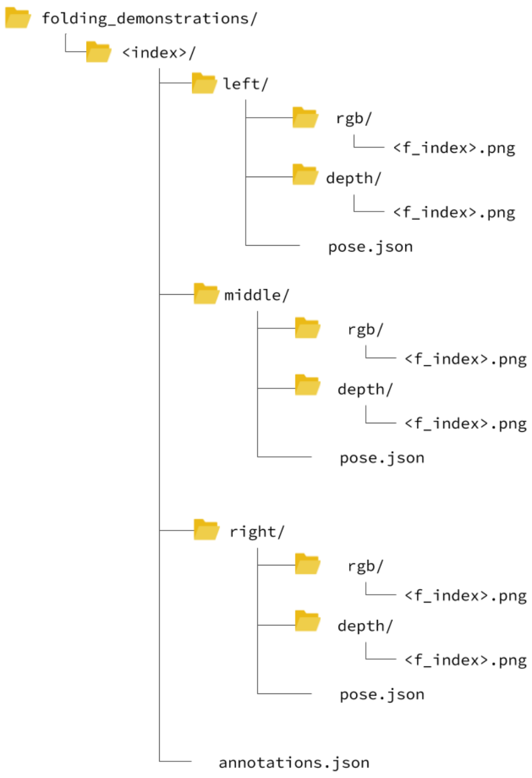


Figure 4. Folder structure of the folding demonstration dataset. There is a folder per folding demonstration, indicated with *<index>*. Each sample contains labelled data in the annotations json file. The images are grouped per perspective and contain rgb and depth images. There are also joint positions available per video perspective.

Quality label	Definition
Follows instructions	The instructions were followed exactly.
Slight variation on instructions	One deviation was made from the instructions.
Very different from instructions	Two or more deviations were made from the instructions.

Table 2. Because not all demonstrators follow the given high-level task instructions, we define a quality label of which the definitions are given in this table.

Data format

The RGB images captured with the Kinect cameras are compressed with the *x265* codec. We use the intrinsic camera calibration parameters to modify the images according to the depth correction. All RGB frames are anonymized by applying colour quantization to the corners of the frame and pasting an ellipsoid colour patch around the face of the demonstrator which was tracked using AlphaPose (Fang et al. 2017).

Each sample in the dataset contains annotations in the *annotations.json* file. The labelled information and data format can be seen in Listing 1. We label which type of textile is being folded and which folding method is being used. We distinguish four categories of folding methods, labelled from *a* to *d*. These categories represent an increasing amount of complexity to learn a certain folding strategy. For example, folding method *a* extensively uses the table to make folds. In contrast, method *b* represents demonstrators making vertical folds while lifting the cloth in the air. Method *c* categorizes folding strategies which requires crossing the hands. Finally, method *d* captures all different strategies not described by the former folding categories, for example rolling up the cloth. All four folding strategies can be used on all types of cloth in the dataset. The different types of textile are labelled as *hoodie*, *shirt* or *towel*. The distribution of the folded clothing is as follows: 88% of the folded clothing are shirts, 9% are towels and 3% hoodies. The timestamp is in `YYYY-MM-DD HH:MM:SS` format. Given that the data is crowdsourced, some variation exist in the way participants followed the high-level process instructions. For example, some demonstrators fold the clothing immediately out of the basket instead of first collecting the pieces on the left side of the table. To indicate to which extent the given process instructions are being followed, we included a quality label in the annotations. This label is useful if consistent, high-quality samples need to be sampled. In the dataset, 86% follow the given instructions, 12% make one deviation while 2% do not follow the given high-level instructions. The exact definitions of the quality label are shown in Table 2.

We labelled each part of the video with a descriptor indicating which step in the folding process the demonstrator is going through. The different steps are named *isolated_grasping*, *unfolding*, *flattening*, *folding* and *stacking*.

We provide human skeleton keypoint trajectories in the file *pose.json*. There are pose trajectories available for each

```
{
    "id": 0,
    "timestamp": "2018-09-30 19:35:06",
    "cloth_type": "hoodie",
    "location_id": 0,
    "nb_frames": 579,
    "folding_method": "a",
    "demonstration_id": 0,
    "nb_folds": {
        "0": 0,
        "66": 1,
        "94": 2,
        "118": 3
    },
    "subtask_changes": {
        "0": "isolated_grasping",
        "42": "unfolding",
        "112": "folding",
        "300": "stacking"
    },
    "quality": "follows instructions"
}
```

Listing 1: annotations.json description

camera perspective per folding sample. The joint positions are stored in the JSON format visible in Listing 2. There is a score associated with each pair of x and y coordinates. This variable, ranging from 0 to 1 indicates the detection confidence that a certain joint is at the given location. In general, the coordinates for every joint positioned beneath the shoulders are less reliable because the subject is sitting on a bench with the legs occluded by the table. We consider this not a problem because the coordinates of the joints of the two arms are reliable and are of importance for the folding task.

```
{
    frame_nr: {
        "LElbow": [x, y, score],
        "RElbow": [x, y, score],
        "LShoulder": [x, y, score],
        "RShoulder": [x, y, score],
        "LWrist": [x, y, score],
        "RWrist": [x, y, score],
        ...
        "confidence": 100.0
    }
}
```

Listing 2: pose.json description

Project website and helper scripts

Along with the data, we provide helper scripts in Python which are available at https://github.com/adverley/folding-demonstrations. The data can be loaded by calling `FoldingDemonstrationDataSet(home_dir)`. We expose the data as a nested dictionary, embedded in a list. This enables an intuitive interface for accessing the data by iterating over the `FoldingDemonstrationDataSet` object and querying specific fields with corresponding key in square brackets. For example, `data[0].annotations['clothing_type']` queries the type of clothing being folded in demonstration 0 while `dataset[42][0]['rgb']['left']` returns the first RGB image of video demonstration 42. A more complete and general-purpose example can be found in Listing 3.

```
from folding_demonstrations.dataset
 ↪  import FoldingDemonstrationDataSet

# Set to the directory where the
 ↪  folding demonstrations dataset is
 ↪  stored
home_dir =
 ↪  '/media/data/folding_data_output'

# Load the data
dataset = FoldingDemonstrationDataSet(⌋
 ↪  home_dir)

# Iterate over data and query
 ↪  available information
for sample in dataset:
    random_frame_nr = 42
    frame = sample[random_frame_nr]
    rgb_left = frame['left']['rgb']
    rgb_middle = frame['middle']['rgb']
    rgb_right = frame['right']['rgb']
    depth_l = frame['left']['depth']
    depth_m = frame['middle']['depth']
    depth_r = frame['right']['depth']
    subtask = frame['subtask']
    reward = frame['reward']
    pose = frame['pose']
```

Listing 3: Example code how to query the dataset

Conclusion

In this paper, we introduce a video dataset with human demonstrations of folding textile, captured via a citizen crowdsourcing project. With this dataset, we aim to fill in a gap in learning deformable objects manipulation, bootstrapped by human examples. We provide 1000 demonstrations with RGB images, depth frames, and joint pose trajectories captured from three perspectives simultaneously. We labelled the data with subtask annotations, folding method, and textile type. Our goal is to provide robotics researchers with a real-world dataset to accelerate the learning from human demonstrations for deformable object manipulation.

Acknowledgements

We would like to thank Jeroen Bourgonjon for helping us organize this project in De Krook, Gertjan Hoste for designing and building the folding table and all participants for providing their folding demonstrations.

This research is supported by the Research Foundation Flanders (FWO) under Grant number 1S54220N.

References

Abbeel P and Ng AY (2004) Apprenticeship learning via inverse reinforcement learning. In: *Proceedings of the twenty-first international conference on Machine learning*. ACM, p. 1.

Doumanoglou A, Stria J, Peleka G, Mariolis I, Petrík V, Kargakos A, Wagner L, Hlaváč V, Kim T and Malassiotis S (2016) Folding clothes autonomously: A complete pipeline. *IEEE Transactions on Robotics* 32(6): 1461–1478. DOI:10.1109/TRO.2016.2602376.

English P, Richardson M and Garzón-Galvis C (2018) From crowdsourcing to extreme citizen science: Participatory research for environmental health. *Annual Review of Public Health* 39(1): 335–350. DOI:10.1146/annurev-publhealth-040617-013702. URL https://doi.org/10.1146/annurev-publhealth-040617-013702. PMID: 29608871.

Fang HS, Xie S, Tai YW and Lu C (2017) RMPE: Regional multi-person pose estimation. In: *ICCV*.

Finn C, Levine S and Abbeel P (2016) Guided cost learning: Deep inverse optimal control via policy optimization. In: *International Conference on Machine Learning*. pp. 49–58.

Fleming P (2019) Robots and organization studies: Why robots might not want to steal your job. *Organization Studies* 40(1): 23–38. DOI:10.1177/0170840618765568. URL https://doi.org/10.1177/0170840618765568.

Foresti GL and Pellegrino FA (2004) Automatic visual recognition of deformable objects for grasping and manipulation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34(3): 325–333. DOI:10.1109/TSMCC.2003.819701.

Levine S, Pastor P, Krizhevsky A, Ibarz J and Quillen D (2018) Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research* 37(4-5): 421–436. DOI:10.1177/0278364917710318. URL https://doi.org/10.1177/0278364917710318.

Liu Z, Luo P, Qiu S, Wang X and Tang X (2016) Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mahler J, Matl M, Satish V, Danielczuk M, DeRose B, McKinley S and Goldberg K (2019) Learning ambidextrous robot grasping policies. *Science Robotics* 4(26). DOI:10.1126/scirobotics.aau4984. URL https://robotics.sciencemag.org/content/4/26/eaau4984.

Maitin-Shepard J, Cusumano-Towner M, Lei J and Abbeel P (2010) Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In: *2010 IEEE International Conference on Robotics and Automation*. pp. 2308–2315. DOI:10.1109/ROBOT.2010.5509439.

Matas J, James S and Davison AJ (2018) Sim-to-real reinforcement learning for deformable object manipulation. In: Billard A, Dragan A, Peters J and Morimoto J (eds.) *Proceedings of The 2nd Conference on Robot Learning*, *Proceedings of Machine Learning Research*, volume 87. PMLR, pp. 734–743. URL http://proceedings.mlr.press/v87/matas18a.html.

Miller S, van den Berg J, Fritz M, Darrell T, Goldberg K and Abbeel P (2012) A geometric approach to robotic laundry folding. *The International Journal of Robotics Research* 31(2): 249–267. DOI:10.1177/0278364911430417. URL https://doi.org/10.1177/0278364911430417.

Seita D, Jamali N, Laskey M, Berenstein R, Tanwani AK, Baskaran P, Iba S, Canny JF and Goldberg K (2018) Robot bed-making: Deep transfer learning using depth sensing of deformable fabric. *CoRR* abs/1809.09810. URL http://arxiv.org/abs/1809.09810.

Sharma P, Mohan L, Pinto L and Gupta A (2018) Multiple interactions made easy (mime): Large scale demonstrations data for imitation. In: *Conference on Robot Learning*. pp. 906–915.

Večerík M, Hester T, Scholz J, Wang F, Pietquin O, Piot B, Heess N, Rothörl T, Lampe T and Riedmiller M (2017) Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817* .

Xiang L, Echtler F, Kerl C, Wiedemeyer T, Lars, hanyazou, Gordon R, Facioni F, laborer2008, Wareham R, Goldhoorn M, alberth, gaborpapp, Fuchs S, jmtatsch, Blake J, Federico, Jungkurth H, Mingze Y, vinouz, Coleman D, Burns B, Rawat R, Mokhov S, Reynolds P, Viau P, Fraissinet-Tachet M, Ludique, Billingham J and Alistair (2016) libfreenect2: Release 0.2. DOI:10.5281/zenodo.50641. URL https://doi.org/10.5281/zenodo.50641.