Learning Grasping Points for Garment Manipulation in Robot-Assisted Dressing

Fan Zhang and Yiannis Demiris

Abstract—Assistive robots have the potential to provide tremendous support for disabled and elderly people in their daily dressing activities. Recent studies on robot-assisted dressing usually simplify the setup of the initial robot configuration by manually attaching the garments on the robot end-effector and positioning them close to the user's arm. A fundamental challenge in automating such a process for robots is computing suitable grasping points on garments that facilitate robotic manipulation. In this paper, we address this problem by introducing a supervised deep neural network to locate a pre-defined grasping point on the garment, using depth images for their invariance to color and texture. To reduce the amount of real data required, which is costly to collect, we leverage the power of simulation to produce large amounts of labeled data. The network is jointly trained with synthetic datasets of depth images and a limited amount of real data. We introduce a robot-assisted dressing system that combines the grasping point prediction method, with a grasping and manipulation strategy which takes grasping orientation computation and robot-garment collision avoidance into account. The experimental results demonstrate that our method is capable of yielding accurate grasping point estimations. The proposed dressing system enables the Baxter robot to autonomously grasp a hospital gown hung on a rail, bring it close to the user and successfully dress the upper-body.

I. INTRODUCTION

Dressing is a basic aspect of the daily life in which many disabled and elderly people require assistance due to impairments. Studies have reported that of all the activities of daily living, dressing has shown the highest burden on caregiving staff and the lowest use of assistive technologies [1, 2]. In this paper, we focus on robot-assisted dressing, which has the potential to alleviate such problems.

Recent studies on robot-assisted dressing usually simplify the setup of the initial robot configuration before dressing sequences by manually attaching the garments on the robot end-effector [3-8]. Although these works successfully enable the robots to provide users with dressing assistance in putting on various types of garments, their experiments are usually initiated with the user's arm already in or close to the sleeve opening. Robots should be endowed with the ability to autonomously recognize the garment state, grasp and hand the garment to the user and subsequently complete the dressing process. However, garment manipulation for robots remains challenging because of their inherent high-dimensionality and non-linear features [9]. Instead of modelling the whole

Fig. 1: The Baxter humanoid robot autonomously detects a grasping point on a hospital gown hung on a rail, grasps and brings the garment close to the human user and subsequently complete the upper-body dressing. This is achieved with a grasping point prediction using a deep neural network, combined with a robotic grasping and manipulation strategy introduced in this paper.

garment state, in this work, we focus on computing suitable garment grasping points that facilitate robotic manipulation and dressing.

We propose using Convolutional Neural Networks (CNN) to estimate the Cartesian positions of a pre-defined grasping point on the garment. Depth images are used as the input modality for their invariance to different colors and textures. Other studies have also investigated the similar problem of detecting garment grasping points using CNN, where either only real depth data of hung garments [9] or only synthetic data [10] is used for network training. However, real-world labeled data, especially in the domain of robotics, is notoriously costly to collect [11]. To reduce the amount of real data required, the power of simulation is usually leveraged to continuously generate large amounts of labeled data. However, training models with synthetic depth images may not readily transfer to real-world applications [12], and thus the prediction accuracy cannot be guaranteed.

In this work, we propose to train the deep neural network with both synthetic and real datasets. The idea is to use as little real-world data as possible, with the help of simulation, to generate accurate grasping points estimation, and the corresponding successful physical robotic grasping. The method is applied in the scenario encountered in the robot-assisted

The authors are with the Personal Robotics Lab, Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2BT, U.K. (e-mail: f.zhang16@imperial.ac.uk; y.demiris@imperial.ac.uk). This research is supported in part by EPSRC Grant EP/S032398/1, and a Royal Academy of Engineering Chair in Emerging Technologies to YD.

dressing applications. As presented in Fig. 1, a grasping point needs to be located on a gown, which is naturally hung on a rail to simulate home or hospital environments.

Having the predicted grasping point positions, we present a grasping strategy for actual robot execution. As pointed out by [10, 13], defective grasping is a non-trivial problem that has caused most of the failures in garment manipulation with robots. Therefore, we calculate the orientation for each grasping point from the garment point cloud. The trajectory of the robot approaching the garment is planned to avoid collisions between the robot and the garment, which may lead to changing the grasping point position before the actual grasping is executed.

Experimental evaluation of the grasping point prediction, and the corresponding physical robotic grasping and dressing, shows that it outperforms three baselines in the literature, including network training using only synthetic data (similar to the vast majority of literature [13]), training only using synthetic data modified by adding simulated depth noise (similar to [10]), and training using only real data (similar to [9]). We also investigate an image-to-image translation approach in our experiments for bridging the domain gaps. This is one of the first attempts to quantitatively and qualitatively report the garment grasping and dressing results for the robot-assisted clothing task.

The main contributions of this paper can be summarized as follows:

- 1) A supervised deep neural network is presented to yield grasping point estimations. To reduce the amount of real data required, we built a simulation framework to produce large amounts of labeled garment depth images. The network is jointly trained using a limited amount of real-world data, with large amounts of simulated data.
- 2) A robot-assisted dressing system is proposed, combining the grasping point prediction method with a grasping and manipulation strategy, taking into account grasping orientation computation and robot-garment collision avoidance.

II. RELATED WORK

A. Robot-Assisted Dressing

There exist several approaches that tackle the robot-assisted dressing task in home environments. For example, reinforcement learning has been successfully used to enable a humanoid robot to wrap a scarf [14] and to put on a T-shirt for mannequins with different postures [15]. Learning by demonstration techniques have been performed to encode dressing skills in [5, 16]. User capabilities and preferences have been studied to personalize dressing assistance for users suffering from different impairments [3, 7]. Force information has been considered in [17] to adjust the robot motions to minimize the interaction force between the robot and the user. User movements have been tracked during the dressing process to plan the dressing trajectory in real-time [8, 18]. While these different studies have led to promising results of dressing, they usually simplify the setup of their experiments by positioning the robot end-effector, with the garment manually attached on, close to the user's arm.

Initial robot configurations before the dressing sequence have been taken into account in recent studies. For example, in [19] a robot approaches the user with a shoe attached on the gripper and successfully inserts the shoe on the user's foot. The robot motions are planned and adapted according to user's speech and movement reactions. [20] has optimized the initial robot configuration considering user's arm postures, so as to enable a robot with a mobile base to approach the user and pull a hospital gown onto the person's upper-body. However, in these studies, garments are still manually attached on the robot end-effector. Robots should be endowed with the ability to autonomously achieve such garment manipulation, *i.e.*, detecting grasping points, grasping and bringing the garment close to the user, and subsequently completing the dressing process.

B. Garment Grasping Point Computation

A significant amount of research has focused on garment identification and manipulation for robots. Early research using computer vision usually extracts handcrafted features, *e.g.*, wrinkles [20, 21], edges and corners [22], shapes [23], and volumetric features [24]. Such features have been widely explored in a variety of contexts, such as robotic ironing [25], laundry folding [26] and sewing [27]. In these research, the garments are usually manually strewn across a flat surface in either a spread or crumpled state, or hold by a robot gripper in midair using gravity to help expose borders for grasping points detection.

Recently, machine learning methods have enjoyed substantial attention in the garment perception community [28]. Random forests techniques have been successfully utilized to learn grasping points on fabrics [29], and [30] followed by proposing a complete pipeline for garment folding. Convolutional Neural Networks (CNN) have been shown to be effective in learning discriminant features from raw depth images, instead of hand-engineered features extraction [31]. Studies [9, 10, 13] have taken advantage of such perception power of CNN and made promising progress in identifying grasping points on garments. As discussed in Section II, neural network training in these studies relies on either only real depth data of hung garments or only synthetic ones. For example, [9] has collected a real dataset consisting of more than 80000 depth images.

Such real labeled data collection is expensive and cumbersome, often requiring days for a single task, which makes it challenging to generalize the method to various types of clothes. Thus in [10] large-scale simulations are leveraged to collect knowledge from synthetic depth images. But with this comes the issue of transferring gained knowledge from simulation to the real world. In [13], simulation-based training results have been further tested on depth images of real garments and achieved a prediction error of approximately 10cm. An interesting question would be how to improve the prediction accuracy with a limited amount of real data, while in parallel taking advantage of the power of simulation. In this paper, we also explore deploying the trained network in the application of robot-assisted dressing, specifically,

Fig. 2: (left) The garment model in the Autodesk Maya physics engine. Cameras are set around the garment to capture depth images from all points of view. For clarity, not all cameras in simulation are shown in this figure. (middle) Simulated garment falling based on gravity. (right) A real gown hung on a rail to simulate home or hospital environments.

picking up a gown hung on a rail which is a scenario usually encountered in home or hospital environment.

III. GARMENT GRASPING POINT DETECTION

In this section, we present a deep neural network to estimate Cartesian positions of a pre-defined grasping point on a hospital gown hung on a rail. Both real and synthetic depth images of the garment are employed in the design and evaluation of the networks.

A. Dataset Acquisition

Real Garment Dataset: As shown in Fig. 2-right, a hospital gown is naturally hung on a rail, while the hanging point is randomly around the collar of the gown. A Kinect camera is placed 100cm in front and 30cm lower than the hanging point. Based on preliminary experiments, we pre-define a grasping point (black markers in Fig. 2), selected to be appropriate for the dressing task. To avoid occlusions, markers are attached on both sides of the garment. We continuously hang and remove the gown from the rail and obtain a real garment dataset consisting of 6000 depth images, as well as the grasping point positions recorded. The whole process takes approximately 40 hours. The acquired depth images are cropped to remove the non-garment part of the image and resized to 160×160 pixels. The garment and the rail are segmented from the background by thresholding the depth between 70cm and 130cm, and the rail is further removed based on its position.

Synthetic Garment Dataset: Synthetic dataset of depth images are acquired using the physics simulator Autodesk Maya [32]. A 3D model of the real gown has been built, as shown in Fig. 2-left. The grasping point has been defined on the simulated garment approximately matching the topology of the real one (orange dots in Fig. 2). The garment is hung randomly from the points near the collar and simulated falling based on gravity is executed (Fig. 2-middle). Although in the real world the camera is only facing one direction of the garment, as the falling and deformation in simulation are highly unpredictable, we put 36 cameras around the garment and capture depth images from all points of view after the garment is stabilized. To imitate the real

Fig. 3: The CNN network architecture. The input is the resized (160×160) depth images, while the output is the Cartesian positions of the grasping point. Convolutional and fully-connected layers are abbreviated with ‘Conv’ and ‘FC’ respectively. The dotted lines show the training phase.

environment, camera parameters (e.g., horizontal and vertical viewing angles, focal length, the distance to the garment) are set the same, and image cropping and garment segmentation are carried out in the same procedure. This process is autonomously repeated and 62000 synthetic depth images are obtained, while retrieving the pre-defined grasping point positions. Note that our method relies on different models for different types of clothing.

B. Deep Neural Network Structure

We apply a convolutional neural network modified from the structure presented in [31], as depicted in Fig. 3. The network directly maps from the spatial relation of pixels in depth images to Cartesian positions where the robot will grasp. Other components of the neural network include: a Mean Squared Error loss function using Adam optimizer, 50% dropout in fully-connected layers, a batch size of 64 and a learning rate of 0.0001.

The network is jointly trained with both synthetic datasets of depth images and real data. Training and testing results will be detailed in Section V-B.

IV. GARMENT GRASPING AND MANIPULATION FOR ROBOT-ASSISTED DRESSING

Having the predicted grasping point positions, in this section we introduce a grasping strategy for robots. We compute the specific orientation for each grasping point from garment point cloud. Collision avoidance between the robot and the garment is considered in the process of the robot approaching the garment, to prevent changing the grasping point position before the actual grasping.

A. Point Cloud Filtering

As the data is indexed according to its spatial distribution, there is a direct correspondence between the pixels of depth images and the point cloud indexes. As described in Section III-A, the garment has been segmented out from the background. Here we segment out the garment point cloud with the same pixel coordinates. We then apply a widely-used voxelized grid filter [33] to considerably reduce the number of points and effectively represent the 3D geometric features of the garment.

Fig. 4: Grasping orientation computation. The surface normal (*i.e.*, Y -axis) at the grasping point can be estimated in either one of two opposite directions, which leads to two possible different frames. The frame is determined by calculating P' , which is regarded as the target point of the robot approaching trajectory. (left) The correct frame so that the robot can approach the garment and stop at point P' . (right) This is an example of wrong frame that leads to the robot reaching point P' and colliding with the garment.

B. Grasping Orientation computation

The pose of the garment is defined differently from the rigid object pose as garments are highly deformable. Instead of directly relating the pose to the overall shape of the garment, it can be defined in relation to the desired grasping point, pointing to the direction that the gripper can approach and grasp the garment ideally. In our case, due to the accurate estimations of the grasping point along the garment edge, the orientation of the point can effectively represent the shape of edge area to facilitate a proper robotic grasping.

Fig. 4-left shows the ideal grasping orientation for robot end-effectors. Specifically, Y -axis should be aligned with the gripper opening/closing direction, while Z -axis aligned with the parallel fingers direction, and X -axis approximately aligned with the garment edge direction.

We propose to compute the orientation for each grasping point from point cloud. Having the predicted grasping point P , we first locate its nearest point $\hat{P}$ in the point cloud. Surface normals are estimated using Open3D library [34]. Then the normal $\vec{n}$ at point $\hat{P}$ is defined as the Y -axis. The vector $\vec{H\hat{P}}$ from the hanging point H to the point $\hat{P}$ is obtained. The cross product of $\vec{H\hat{P}}$ and $\vec{n}$ is computed and regarded as Z -axis. X -axis is obtained accordingly by computing the cross product of Z -axis and Y -axis. Then this frame is defined as the local frame of the robot end-effector.

Note that surface normal estimation (in our case, the point normal Y -axis) can generate either one of two vectors that are in opposite directions. This leads to two possible frames, as shown in Fig. 4-left and Fig. 4-right. To determine the Y -axis direction, we first calculate points P' based on the two frames, which are 5cm way from $\hat{P}$ in the opposite Z -axis direction. Point P' is regarded as the target point of the robot approaching trajectory before grasping the garment. Then we respectively compute the average distance between point P' and its k -nearest neighbors in the point cloud in these two frames. Point P' should be out of the point cloud (Fig. 4-left) to prevent robot-garment collisions. Therefore, the point with the larger average distance is adopted and thus the frame is determined. Then the local gripper coordinate

is transformed to the global Baxter robot frame F_B .

C. Garment Manipulation

The robot approaching trajectory is designed as moving the robot gripper from the initial state to point P' , which is 5cm way from the grasping point $\hat{P}$ in the opposite Z -axis direction as described in Section IV-B. The robot motion is planned using Moveit framework [35]. Garment point cloud from the Kinect sensor is integrated with Moveit using OctoMap, which prevents the robot-garment collisions during the process of the robot gripper approaching the garment. After reaching point P' , the robot continues its action to point $\hat{P}$ and grasps the garment.

D. Robot-Assisted Dressing

After the robot grasps the garment, the robot is programmed to move 10cm along the Y -axis of the global coordinate F_B to remove the garment from the rail. The human user postures, including the user's hand, elbow, and shoulder positions, are detected by the depth camera and defined as the waypoints of the dressing trajectory. Then, the robot gripper starts to move the clothes to the initial hand position and subsequently executes the dressing trajectory. Fig. 5 presents an example of the whole pipeline, which shows that the robot autonomously detects the grasping point of a gown hung on a rail, picks up the gown, brings it close to the user and completes the upper-body dressing.

V. EXPERIMENTS AND RESULTS

We conduct two experiments to investigate: 1) the accuracy of our proposed grasping point prediction method (Grasping Point Prediction Evaluation); and 2) the quality of the whole proposed dressing system (Robot-Assisted Dressing Evaluation). A supplementary video also illustrates the performance of the proposed method¹.

A. Experimental Setup

We implement the proposed method on the Baxter humanoid robot. As described in Section III-A and shown in Fig. 5 a gown is naturally hung on a rail to imitate home or hospital environments. A Kinect camera is placed 100cm in front and 30cm lower than the hanging point. User postures are detected with another ASUS Xtion camera using HRNet library [36]. The coordinate transformation between the cameras and the robot has been determined prior to the experiments.

B. Grasping Point Prediction Evaluation

In the first experiment, we investigate the accuracy of our proposed grasping point prediction method and explore the cross-domain knowledge transfer between the synthetic and real datasets. As presented in Section III-A, we have acquired a real dataset consisting of 6000 depth images. These are divided into training (5000 images), validation (500 images) and test (500 images) sets. We train the network with all

¹<http://www.imperial.ac.uk/personal-robotics/videos/>

Fig. 5: Snapshots of the robot-assisted dressing sequence. (a) Grasping point is predicted and orientation is computed. (b) Robot approaches the garment. (c) Robot grasps the garment. (d) Robot brings the garment close to the user. (e) Robot completes dressing the right arm.

Fig. 6: The boxplot pictures the grasping point prediction errors of the test sets with 5-fold cross validation. The central dot corresponds to the median value of the errors, while the sides of the box refer to the first and third quartiles of the data. The whiskers extend to the most extreme data points that are not considered as outliers, which are displayed as individual circles. The x-axis corresponds to the different amounts of real data (*i.e.*, 0, 100, 500, 1000, 2000 and 5000) used in network training, as described in Section V-B. In column 1, zero real image implies that all training is done in simulation, and thus results of training with real images only and CycleGAN are not available. Best viewed in color.

synthetic images, together with 0, 100, 500, 1000, 2000 and 5000 real images respectively. For all the training in this experiment, the validation sets are 500 real images, and the test sets are another 500 real images. Due to the small size of the real dataset, 5-fold cross validation is employed.

We benchmark our results versus four baselines:

1) Network training with only synthetic datasets: This baseline corresponds to the approach using only synthetic data, as customary in the literature [13].

2) Network training with only synthetic datasets modified by adding simulated depth noise: In simulation, synthetic images are devoid of any noise, and thus the garment borders appear very clear. In this baseline, we apply a simulated depth noise model proposed in [37], which uses random offsets to shift pixel locations and adds Gaussian noise. This baseline corresponds to the method used in literature [10].

3) Network training with only real datasets: This baseline is adopted by [9] to train the network using only real images.

4) Network training with simulated datasets modified by an image-to-image translation approach, together with real images: In this baseline, we investigate on using a state-of-the-art image-to-image translation method called Cycle-Consistent Adversarial Network (CycleGAN) [38] to synthesize realistically looking ‘sim2real’ images. CycleGAN has

been widely explored to bridge the domain gap, including, robotics visuomotor learning [11] (RGB to RGB domain), depth estimation [39] (RGB to depth domain), and depth noise prediction [40] (depth to depth domain). Here the CycleGAN is trained with all synthetic images and 100, 500, 1000, 2000 and 5000 real images respectively.

Fig. 6 illustrates the results of grasping point prediction errors of the test sets with 5-fold cross validation. Three conclusions can be made from Fig. 6

1) The prediction error when training the network with only synthetic images (10.67cm, column 1, gray, baseline 1) is 2.88, 4.10, 5.01, 6.27, 7.82 times larger than the errors when jointly training with synthetic images and 100, 500, 1000, 2000, 5000 real images respectively (column 2-6, gray). Similar results can be obtained when analyzing the prediction error from the network training using only synthetic images modified by adding simulated depth noise (column 1, orange, baseline 2). This indicates that jointly training the network using synthetic data, with a limited amount of real data, can significantly improve the prediction accuracy against training with only synthetic data.

2) The prediction errors when training the network using only 100, 500, 1000, 2000, 5000 real images (column 2-6, red, baseline 3) are 1.43, 1.31, 1.20, 1.04, 1.02 times larger than the errors when jointly training with synthetic images and the corresponding amount of real images respectively (column 2-6, gray). This indicates that jointly training the network using real data, with the help of large amounts of simulated data, can also improve the prediction accuracy than training with only corresponding amount of real data.

We can also see that synthetic data significantly improves prediction performance when very small amounts of real data is available (*i.e.*, 100 and 500 real images). For the cases of 1000, 2000 and 5000 real images, only small improvements have been obtained. This shows that more gained knowledge from simulation has been transferred to the real world when real-world data is insufficient.

3) The prediction errors when jointly training the network with simulated datasets modified by CycleGAN, together with real images (column 2-6, blue, baseline 4) are all within 5% difference with the errors when jointly training with the original synthetic images and the corresponding amount of real images respectively (column 2-6, gray). This demonstrates a weak gap reducing between two domains using CycleGAN in our context.

C. Robot-Assisted Dressing Evaluation

In the second experiment, we deploy the trained network on the physical robot to evaluate grasping and dressing performance. As shown in Section V-B, adding simulated Kinect noise and transferring using CyclGAN have not significantly improved the prediction accuracy, compared to using the original synthetic images. Therefore, in this experiment we test the models obtained by training original synthetic images on the real robot. Overall, 11 training models are deployed, including 6 models of training with both synthetic and varying amounts of real data (*i.e.*, 0, 100, 500, 1000, 2000, 5000), and 5 models of training with only corresponding amount of real data. For each training model, 50 replications of trials have been carried out. For each trial, a grasping point is predicted using the training model given a new depth image, and subsequently the robot grasps the point and executes the dressing task. Fig. 7 shows the success rates of garment grasping and dressing.

We first focus on the grasping success rates when 0, 100, 500 real images are involved in training respectively. Fig. 7 shows that training only on synthetic data achieves a low success rate (6%), which indicates the existence of the reality gap. When jointly training the network using synthetic images, with 100 and 500 real data, the grasping success rapidly increases to 46% and 78% respectively against simulation-only training. Meanwhile, 24% and 22% improvements have been respectively observed when comparing joint network training using 100 and 500 real data, against training only with corresponding amount of real data.

Fig. 7 also presents that when 1000, 2000 and 5000 real images are involved in training respectively, almost equal grasping success rates are achieved when comparing jointly network training and real-only training. Fig. 6 shows that when the amount of real data reaches 1000, the prediction error falls within approximately 2.5cm. Each time having a predicted point, we always first locate its nearest point in the point cloud. Therefore, such equal performances could be explained that when the prediction error is small, points that are located in the point cloud could be close to each other, and thus comparable grasping performances are obtained.

We can see from Fig. 7 that the same conclusion as Fig. 6 can be drawn: synthetic data significantly improves grasping success rates when very small amounts of real data is available (*i.e.*, 100 and 500 real images). This success highlights the efficient knowledge transfer from simulation to real world when real-world data is insufficient. Similar conclusions can be made when analyzing dressing performance.

Another interesting observation is that network training using synthetic data with another 500 real images attains around 78% grasping success, which is a comparable performance to network training using 1000 real images only, resulting in a reduction of real-world data by 50%. Almost equal dressing success rates are also achieved using such two trained models.

We also analyze the failure cases of grasping and dressing. Besides the grasping point prediction error, we have observed

Fig. 7: Success rates of grasping and dressing. The x-axis corresponds to the varying amounts of real data. The panel shows that synthetic data significantly improves grasping and dressing success rates when very small amounts of real data is available (*i.e.*, 100 and 500 real images). Best viewed in color.

that another main cause of grasping failures is the inaccurate point normal estimation, which leads to a defective grasping orientation, especially in some cases when the garment edge area is in vertical configurations to the camera. As discussed in Section IV-B, we use the orientation of the grasping point close to the edge to represent the orientation of the overall edge area. With an inaccurate orientation, the parallel robot grippers may not approach garment along its edge direction but collide with it. Fig. 7 also shows that dressing success rate is usually slightly lower than grasping success rate. In some cases of our experiments, when the garment edge is close to or occluded by the other parts of the garment, the robot grasps multiple layers, leading to the occlusions of the sleeve opening. We observe that the garment can tangle itself in the process of the robot approaching the user, which also makes the sleeve opening not easily accessible for the user.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a framework to automate the garment manipulation process for robot-assisted dressing. A deep neural network was presented to locate a pre-defined grasping point on a garment, trained using both synthetic data of garment depth images and real-world data. A robot-assisted dressing system was introduced that combines the grasping point prediction method, with a grasping and manipulation strategy which takes grasping orientation computation and robot-garment collision avoidance into account.

Future extensions include using active vision (*e.g.*, [41]), with the help of a Baxter robot moving base, to obtain multiple views of the garment and generate a more robust prediction. We also plan to investigate on autonomously choosing the grasping point on the garment instead of pre-defining it using reinforcement learning, which has the potential to tackle the failure cases when the predicted grasping point is occluded.

ACKNOWLEDGEMENTS

We are grateful to the Personal Robotics Laboratory members for fruitful discussions and to Dr Antoine Cully for his assistance with the setup of the simulation environment.

REFERENCES

- [1] T. L. Mitzner, T. L. Chen, C. C. Kemp, and W. A. Rogers, "Identifying the potential for robotics to assist older adults in different living environments," *International journal of social robotics*, vol. 6, no. 2, pp. 213–227, 2014.
- [2] B. J. Dudgeon, J. M. Hoffman, M. A. Ciol, A. Shumway-Cook, K. M. Yorkston, and L. Chan, "Managing activity difficulties at home: a survey of medicare beneficiaries," *Archives of physical medicine and rehabilitation*, vol. 89, no. 7, pp. 1256–1261, 2008.
- [3] F. Zhang, A. Cully, and Y. Demiris, "Personalized robot-assisted dressing using user modeling in latent spaces," in *Intelligent Robots and Systems, IEEE/RSJ International Conference on*. IEEE, 2017, pp. 3610–3603.
- [4] Z. Erickson, M. Collier, A. Kapusta, and C. C. Kemp, "Tracking human pose during robot-assisted dressing using single-axis capacitive proximity sensing," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2245–2252, 2018.
- [5] E. Pignat and S. Calinon, "Learning adaptive dressing assistance from human demonstration," *Robotics and Autonomous Systems*, vol. 93, pp. 61–75, 2017.
- [6] N. Koganti, T. Tamei, K. Ikeda, and T. Shibata, "Bayesian nonparametric learning of cloth models for real-time state estimation," *IEEE Transactions on Robotics*, 2017.
- [7] Y. Gao, H. J. Chang, and Y. Demiris, "User modelling for personalised dressing assistance by humanoid robots," in *Intelligent Robots and Systems, IEEE/RSJ International Conference on*. IEEE, 2015, pp. 1840–1845.
- [8] F. Zhang, A. Cully, and Y. Demiris, "Probabilistic real-time user posture tracking for personalized robot-assisted dressing," *IEEE Transactions on Robotics*, 2019.
- [9] I. Mariolis, G. Peleka, A. Kargakos, and S. Malassiotis, "Pose and category recognition of highly deformable objects using deep learning," in *2015 International Conference on Advanced Robotics (ICAR)*. IEEE, 2015, pp. 655–662.
- [10] 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.
- [11] P. D. Nguyen, T. Fischer, H. J. Chang, U. Pattacini, G. Metta, and Y. Demiris, "Transferring visuomotor learning from simulation to the real world for robotics manipulation tasks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 6667–6674.
- [12] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 627–12 637.
- [13] K. Saxena and T. Shibata, "Garment recognition and grasping point detection for clothing assistance task using deep learning," in *2019 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2019, pp. 632–637.
- [14] A. Colomé, A. Planells, and C. Torras, "A friction-model-based framework for reinforcement learning of robotic tasks in non-rigid environments," in *IEEE International Conference on Robotics and Automation*. IEEE, 2015, pp. 5649–5654.
- [15] T. Matsubara, D. Shinohara, and M. Kidode, "Reinforcement learning of a motor skill for wearing a t-shirt using topology coordinates," *Advanced Robotics*, vol. 27, no. 7, pp. 513–524, 2013.
- [16] G. Canal, G. Alenya, and C. Torras, "Personalization framework for adaptive robotic feeding assistance," in *International Conference on Social Robotics*. Springer, 2016, pp. 22–31.
- [17] Z. Erickson, A. Clegg, W. Yu, G. Turk, C. K. Liu, and C. C. Kemp, "What does the person feel? learning to infer applied forces during robot-assisted dressing," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6058–6065.
- [18] Z. Erickson, H. M. Clever, V. Gangaram, G. Turk, C. K. Liu, and C. C. Kemp, "Multidimensional capacitive sensing for robot-assisted dressing and bathing," *arXiv preprint arXiv:1904.02111*, 2019.
- [19] G. Canal, G. Alenya, and C. Torras, "Adapting robot task planning to user preferences: an assistive shoe dressing example," *Autonomous Robots*, vol. 43, no. 6, pp. 1343–1356, 2019.
- [20] A. Kapusta, Z. Erickson, H. M. Clever, W. Yu, C. K. Liu, G. Turk, and C. C. Kemp, "Personalized collaborative plans for robot-assisted dressing via optimization and simulation," *Autonomous Robots*, pp. 1–25, 2019.
- [21] L. M. Martínez and J. Ruiz-del Solar, "Recognition of grasp points for clothes manipulation under unconstrained conditions," in *Robot World Cup*. Springer, 2017, pp. 350–362.
- [22] C. Kampouris, I. Mariolis, G. Peleka, E. Skartados, A. Kargakos, D. Triantafyllou, and S. Malassiotis, "Multi-sensorial and explorative recognition of garments and their material properties in unconstrained environment," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1656–1663.
- [23] M. Cusumano-Towner, A. Singh, S. Miller, J. F. O'Brien, and P. Abbeel, "Bringing clothing into desired configurations with limited perception," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 3893–3900.
- [24] 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)*. IEEE, 2015, pp. 6000–6006.
- [25] Y. Li, X. Hu, D. Xu, Y. Yue, E. Grinspun, and P. K. Allen, "Multi-sensor surface analysis for robotic ironing," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 5670–5676.
- [26] D. F. Gomes, S. Luo, and L. F. Teixeira, "Garmnet: Improving global with local perception for robotic laundry folding," in *Annual Conference Towards Autonomous Robotic Systems*. Springer, 2019, pp. 49–61.
- [27] J. Schrimpf and L. E. Wetterwald, "Experiments towards automated sewing with a multi-robot system," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 5258–5263.
- [28] D. Seita, N. Jamali, M. Laskey, R. Berenstein, A. K. Tanwani, P. Baskaran, S. Iba, J. Canny, and K. Goldberg, "Robot bed-making: Deep transfer learning using depth sensing of deformable fabric," *arXiv preprint arXiv:1809.09810*, 2018.
- [29] A. Doumanoglou, T.-K. Kim, X. Zhao, and S. Malassiotis, "Active random forests: An application to autonomous unfolding of clothes," in *European Conference on Computer Vision*. Springer, 2014, pp. 644–658.
- [30] A. Doumanoglou, J. Stria, G. Peleka, I. Mariolis, V. Petrik, A. Kargakos, L. Wagner, V. Hlaváč, T.-K. Kim, and S. Malassiotis, "Folding clothes autonomously: A complete pipeline," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1461–1478, 2016.
- [31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [32] "Maya," <https://www.autodesk.co.uk/products/maya/overview>, accessed: 2019-09-15.
- [33] M. Arduengo, C. Torras, and L. Sentis, "A versatile framework for robust and adaptive door operation with a mobile manipulator robot," *arXiv preprint arXiv:1902.09051*, 2019.
- [34] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.
- [35] S. Chitta, "Moveit!: an introduction," in *Robot Operating System (ROS)*. Springer, 2016, pp. 3–27.
- [36] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *arXiv preprint arXiv:1902.09212*, 2019.
- [37] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for rgb-d visual odometry, 3d reconstruction and slam," in *2014 IEEE international conference on Robotics and automation (ICRA)*. IEEE, 2014, pp. 1524–1531.
- [38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [39] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9788–9798.
- [40] C. Sweeney, G. Izatt, and R. Tedrake, "A supervised approach to predicting noise in depth images," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 796–802.
- [41] A. Rasouli, P. Lanillos, G. Cheng, and J. K. Tsotsos, "Attention-based active visual search for mobile robots," *arXiv preprint arXiv:1807.10744*, 2018.