Grasping points detection of garments based on deep learning

Chunhua Shan¹, Yongping Xie²
¹ School of Information & Communication Engineering, Dalian University of Technology, Dalian, Liaoning, 116024, China
² School of Information & Communication Engineering, Dalian University of Technology, Dalian, Liaoning, 116024, China
¹shanchunhua@mail.dlut.edu.cn ²xieyp@mail.dlut.edu.cn

Abstract. Robotic manipulation for rigid objects is a relatively easy task, while grasping highly deformable objects, such as garments, is still a big challenge for robots. This research detects the grasping points of hanging garment using deep learning to facilitate robot manipulation. A neural network is proposed to predict the cartesian coordinates and visibility of predefined grasping points. In order to reduce the impact of different clothing colors, depth images are used for the input of the model. It is inaccurate and unrealistic to label data manually because of complex dynamics of clothes; therefore, synthetic dataset is leveraged to train the neural network. This paper makes use of generative adversarial network (GAN) to translate synthetic data to real data. A real dataset is acquired taken by Azure Kinect sensor as the test dataset. The experimental results indicate that our method can provide accurate prediction of the grasping points and can be applied to real scenarios.

1. Introduction
In modern times, robots play an important role in our daily lives. They perform well in some simple tasks, but perform poorly in complex scenarios. In countries with an aging population, the demand for assistive robots has increased. We hope that robots can help elderly people do more than just sweeping the floor and assisting with voice communication. Folding clothes and dressing are the most commonly required but complex operations for robots, because clothing has complex dynamic properties and high degrees of freedom. Robots also lack human advanced perception capabilities such as sensitive touch and vision, as well as the ability to analyse and deal with complex problems. Garments are generally in an unpredictable state, so how to bring a piece of garment to a fixed pose is a critical problem. Before robots perform operations such as drying and folding clothes, placing the clothes in a fixed state is a very important prerequisite. Our project mainly focuses on detecting the grasping points of garments hanging from a single point, so that the robot can grasp the clothes and bring the clothes to a known configuration.

In recent years, deep learning has become an effective tool in many scientific fields due to its powerful ability of reasoning and rapid information processing, such as image recognition. In our research, we use a convolutional neural network (CNN) to predict the cartesian coordinates and visibility of the garment grasping points. Depth images are used as the input to reduce the effects of varying colors and textures of the garments. The cost of manually labelling three-dimensional (3D) data is enormous due to the complex shape and texture of clothing. Therefore, we use synthetic data generated by a physical simulator - MAYA. However, a model trained by synthetic data cannot be directly applied to the real scenario effectively. It is necessary to narrow the gap between the synthetic data and the real
data. A GAN network is used to address the problem in our research. Finally, a real dataset is acquired by an Azure Kinect sensor to evaluate our method. The contributions of our paper are as follows:

1) A deep learning network is proposed to predict the cartesian coordinates and visibility of the clothing grasping points, which can be used in the clothing grabbing situation of the monocular depth image;

2) A large number of simulation datasets are acquired to train our model and translate synthetic data to “real” data using a GAN, and the results shows that our method can be applied to real life scenarios.

2. Related work
Grasping garments is known as a difficult problem in many studies. Robots usually perform different actions according to the type of clothing. M Kaneko’s work [1] has achieved accurate isolation of a piece of clothing from a pile, which is of great significance to subsequent research work. A standard procedure of recognizing the type of garment first and then detecting the predefined grasping points has proposed in [2]. In terms of garment classification, L Sun et al. [3] has an accuracy of 83.2% for recognizing different types of garments. Mariolis et al. [4] has correctly identified garment types 76% in the single view, and 96% in the multi-view. The accuracy of identifying certain types of clothing is even higher provided enough distance. However, grasping points detection for a certain garment is still an awkward problem.

Early research [5]-[8] extracts traditional handcrafted features to locate predefined points. Y. Kita et al. [9] proposed a deformable model-driven method that uses 3D observation data to recognize the state of a hanging garment. This method uses a physical simulator to simulate the hanging data of the clothes, and recognizes the state of the clothes by comparing the observation data with the candidate shapes predicted in advance. With the development of deep neural networks, A Krizhevsky et al. [10] has proven that neural networks are very effective in image feature extraction. Many past research projects [2], [4], [11] use neural networks to extract features and calculate predefined points of grasping. Since neural networks require a lot of training data, they all use simulated data as the input of the neural network. F Zhang et al. [11] use a small amount of real images to improve the detection accuracy and mention a method for data enhancement using GANs. The difference is that we fix the clothes on the end of the manipulator and the situation is more complicated. At the same time, our method predicts the visibility of the grasping points for the monocular environment.

3. Proposed Method
In this section, we present a real-time system to detect the grasping points of garment hanging on the robot end-effector. The overview of the method is shown in figure 1. First, large amounts of synthetic images with label information and real images are acquired by MAYA. Then we train the GAN network to generate the “real” data (generated data) provided the synthetic and real depth images. Finally, the generated data is used to train the neural network which simultaneously predicts the cartesian coordinates and visibility of grasping points. In this study, we focus on detecting two points on the shoulder of a t-shirt, but our method can be easily transferred to different types of garments.
3.1 Dataset Acquisition

Real garment dataset: The garment is hanging from a single point to simulate the state of being grasped by a robot. The Azure Kinect sensor is placed 60 cm to the garment, approximately the length of the robotic arm. During the whole process, the garments are continuously removed and hung up, and 1000 pictures are acquired, during which the depth threshold of the camera is set to 50-80 cm. It is worth mentioning that because only the depth image is used to manually mark the coordinate position of the predefined point, there is a lot of randomness. We mark the clothes in advance and take the color picture together to observe the position of the predefined points. Then calculate the spatial coordinates of the predefined points according to the camera's hardware parameters.

![Figure 2. Environment in the physics simulator MAYA](image)

Synthetic dataset: To acquire a large number of depth images with labels, we use the physical simulator - MAYA[12] for acquisition of simulated data. As shown in figure 2, we create a real 3D model of a t-shirt and predefine the grasping points on both sides of the shoulder. 121 vertices are defined as suspension points on the model to let the model fall naturally by gravity. 36 cameras are set up around the model to obtain depth images from different views. The simulator tracks the predefined points through time and obtains the depth image and label information automatically when the garment is stable. We project the points on the depth image, and compare the distance from camera and the depth on the image to acquire a visibility label. In the whole process, we set different stretch-resistance and bend-resistance parameters and obtained 65,880 pictures with predefined grasping positions.

Generated dataset: In this paper, a GAN network is used to translate synthetic images to “real” images (Generated data). In this task, it is difficult to access paired training data for the GAN due to the complicated configuration of the garment. We apply the network presented by [13], with the input being synthetic images and real images acquired in advance. After training, we can use the generator to generate "real" pictures. We will use real datasets to evaluate the generated dataset in section 4.

![Figure 3. The CNN network architecture of grasping point detection.](image)

3.2 Deep learning model

We propose a deep learning framework to simultaneously predict the spatial location and visibility of grasping points. As shown in figure 3, the model is a CNN, including five convolutional layers and three maximum pooling layers, and the activation function is the ReLU function. The input is a 224×224 depth image and output has three branches which contain the visibility vectors (2,2) and the 3D coordinates (6) of the two grasping points. For the prediction of key point visibility, we treat the prediction task as a binary classification problem instead of a multi-label classification problem, which
avoids the problem of gradient explosion caused by the sigmoid activation function. The loss function of the CNN network is:

\[
\text{Loss}(t) = \omega_1(t) \sum_{i=1}^{2} \mathcal{L}_{\text{cls}}(t) + \omega_2(t) \mathcal{L}_{\text{reg}}(t),
\]

where the first term of the function is standard cross entropy loss, and the second terms is MSE loss which represents the error between prediction and the ground truth.

In order to balance the training speed of classification and detection tasks, we adopt the method presented in [14]. The two variates \(\omega_1\) and \(\omega_2\) share the parameters of the last layer of the network and the initial values are set to 1 and 10 respectively. For the other hyper-parameters, the learning rate is set to 0.001, the batch size is 64, and the number of iterations is 100k.

4. Experiments

In this section, we investigate the applicability of our proposed method, that is, the results of classification and key point detection, in a real scenario. Three types of data are prepared, including the synthetic dataset divided into a training set (52,704 images) and a test set (13,176 images), the generated data with 65,880 images corresponding to the synthetic images and 1000 real images as the real test set. In the ablation experiments, the hyperparameters of the CNN is the same as in the previous section. We use the average absolute error and classification accuracy to evaluate the detection and classification tasks respectively.

The first and second rows in Table 1 show the results trained with simulated data of our network mentioned in Section 3.2. When the synthetic data sets are used to test in the actual scene (second row), almost all the errors of the predicted points are within the range of 20cm, which demonstrates that applying the power of synthetic data is an effective solution to the problem of detecting key points. However, the average absolute error increases by about 5.86cm compared with testing on the real data set, which is caused by the gap between the simulated environment and the actual environment. We use the GAN network to reduce the difference between the two domains (third row). It can be shown that the mean error of the model trained with generated data decreased by 3.28cm compared with that trained with synthetic data. The prediction accuracy has improved compared with synthetic data. It indicates that our proposed method is improves the accuracy of grasping.

Table 1. The percentage of predicted grasping point to ground truth and the mean average error in the different data sets.

| Train Dataset | Test Dataset | 0-2cm | 2-4cm | 4-8cm | 8-12cm | 12-16cm | 16-20cm | Mean Error(cm) |
|---------------|--------------|-------|-------|-------|--------|---------|---------|---------------|
| Synthetic Data | Synthetic Data | 0.340 | 0.453 | 0.695 | 0.9992 | 0.996 | 1.0 | 5.46 |
| Synthetic Data | Real Data | 0.159 | 0.304 | 0.413 | 0.65 | 0.831 | 0.954 | 11.32 |
| Generated Data | Real Data | 0.197 | 0.367 | 0.540 | 0.894 | 0.971 | 0.985 | 8.04 |

Table 2 shows the classification performance of our model. In the real dataset, all the points are visible. The accuracy of using Synthetic data and real data for model training is roughly the same. The classification accuracy is approximately 90%. The high classification accuracy can simplify the path planning of the robotic arm and avoid collisions with garments.

Table 2. The classification accuracy and average distance to ground truth in the real test data.

| Train Dataset | Test Dataset | Classification Accuracy |
|---------------|--------------|-------------------------|
| Synthetic Data | Real Data | 0.89 |
| Generated Data | Real Data | 0.91 |

5. Conclusion and future work

In this paper, we propose a deep learning framework to predict the position and visibility of grasping highly deformable objects, and translate synthetic data to real data using GAN network. Our method can
be applied to the garment grasping task based on monocular depth cameras. For the robotic gripping task, the cooperation of the software and hardware environment is required. Any problem in any link will cause the grasping operation to fail. Based on our current work, our future research will tend to the orientation of reference grasp points and robot path planning.

Acknowledgments
This research is a sub-project of the automatic clothing folding project of the National Supercomputing Center in Wuxi. We are grateful for their assistance.

References
[1] Kaneko, M., Kakikura, M. (2001) Planning strategy for putting away laundry-isolating and unfolding task. In: Proceedings of the 2001 IEEE International Symposium on Assembly and Task Planning (ISATP2001). Fukuoka, Japan. pp. 429-434.
[2] Corona, Enric, Guillem, A., Antonio, G., Carme, T. (2018) Active garment recognition and target grasping point detection using deep learning. Pattern Recognition, 74: 629-641
[3] Sun, L., Aragon-Camarasa, G., Rogers, S., Stolkin, R., Siebert, J. P. (2017) 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 (IROS). Vancouver, BC. pp. 6699-6706.
[4] Mariolis, Ioannis, Georgia, P., Andreas, K., Sotiris, M. (2015) Pose and category recognition of highly deformable objects using deep learning. In: In Advanced Robotics (ICAR). Istanbul, Turkey. pp. 655-662.
[5] Kapusta, A., Erickson, Z., Clever, H. M., Yu, W., Liu, C. K., Turk, G., Kemp, C. C. (2019) Personalized collaborative plans for robot-assisted dressing via optimization and simulation. Autonomous Robots, 43: 1–25.
[6] Martinez, L. M., Ruiz-del Solar, J. (2017) Recognition of grasp points for clothes manipulation under unconstrained conditions. In: Robot World Cup. Springer. pp. 350–362.
[7] Cusumano-Towner, M., Singh, A., Miller, S., O’Brien, J. F., Abbeel, P. (2011) Bringing clothing into desired configurations with limited perception. In: 2011 IEEE International Conference on Robotics and Automation. Shanghai, China. pp. 3893–3900.
[8] Li, Y., Yue, Y., Xu, D., Grinspun, E., Allen, P. K. (2015) Folding deformable objects using predictive simulation and trajectory optimization. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg, Germany. pp. 6000–6006.
[9] Kita, Y., Ueshiba, T., Neo, E. S., Kita, N. (2009) Clothes state recognition using 3D observed data. In: 2009 IEEE International Conference on Robotics and Automation, Kobe. pp. 1220-1225.
[10] Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012) Imagenet classification with deep convolutional neural networks. In: Conference on Neural Information Processing Systems (NIPS). Harrah's and Harvey's, Lake Tahoe. pp. 1097–1105.
[11] Zhang, F., Demiris, Y. (2020) Learning Grasping Points for Garment Manipulation in Robot-Assisted Dressing. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France. pp. 9114-9120.
[12] “Maya,” https://www.autodesk.co.uk/products/maya/overview, accessed: 2019-09-15
[13] Mejjati Y., Christian, R., James, T., Darren, C. (2018) Unsupervised Attention-guided Image-to-Image Translation. In: Conference on Neural Information Processing Systems (NIPS), Montréal, Canada, pp. 3693–3703.
[14] Zhao, C., Vijay, B., Lee, C., Rabinovich, A. (2018) GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In: Proceedings of the 35th International Conference on Machine Learning (2018). Stockholm, Sweden. pp. 793-802.