Chapter

Visual-Tactile Fusion for Robotic Stable Grasping

Bin Fang, Chao Yang, Fuchun Sun and Huaping Liu

Abstract

The stable grasp is the basis of robotic manipulation. It requires balance of the contact forces and the operated object. The status of the grasp determined by vision is direct according to the object's shape or texture, but quite challenging. The tactile sensor can provide the effective way. In this work, we propose the visual-tactile fusion framework for predicting the grasp. Meanwhile, the object intrinsic property is also used. More than 2550 grasping trials using a novel robot hand with multiple tactile sensors are collected. And visual-tactile intrinsic deep neural network (DNN) is evaluated to prove the performance. The experimental results show the superiority of the proposed method.

Keywords: stable, grasp, tactile, visual, deep neural network

1. Introduction

In recent years, dexterous robotic manipulation increasingly attracts worldwide attention, because it plays an important role in robotic service. Furthermore, the stable grasp is the basis of manipulation. However, stable grasp is still challenging, since it depends on various factors, such as the actuator, sensor, movement, object, environment, etc. With the development of the neural network, the data-driven methods [1] become popular. For example, Levine et al. used 14 robots to randomly grasp over 800,000 times for collecting the data and training the convolutional neural network (CNN) [2]. Guo et al. trained the deep neural network (DNN) with 12 K-labeled images to learn the end-to-end grasping polices [3]. Mahler et al. built the dataset that included millions of point cloud data for training Grasp Quality Convolutional Neural Network (GQ-CNN) with an analytic metric. Then GQ-CNN developed the optimal grasp strategy that achieves 93% success rate for eight kinds of objects [4–6]. Zhang et al. trained robots to manipulate objects by videos that were made by virtual reality (VR). For pick-place tasks, the success rate was increased when the number of samples increased [7]. Therefore, sufficient high-quality data is important for robotic grasping.

Nowadays, a few datasets of robot grasping have been developed. Playpen dataset obtains 60-hour grasping data of robot PR2 with RGBD cameras [8]. Columbia dataset collects about 22,000 grasping samples via the GraspIt! simulator [9]. Besides experiments with robots and numerical simulations, human manipulation videos are also useful. Self-supervised learning algorithms are developed from demonstration of videos [10]. While the above datasets focus on the whole grasping process, there are other datasets that concentrate on specific tasks, like grasp planning and slip detection. Pinto et al. instructed robots to automatically generate labeled images for grasp planning with 50,000 times by self-supervised learning algorithms [11]. MIT built the grasp dataset by vision-based tactile sensor and external vision [12]. While some experiments produced slip with extra force or fix objects [13, 14], researchers recorded the actual random grasping process with 46% failure results in 1000 times grasp [15, 16]. The real data can contribute to the precision grasping [17]. In daily life overabundance of the object's types leads to the difficulty of building datasets. Some researchers select the common objects and build 3D object set models such as KIT objects [18], YCB object set [19], etc. They are more convenient for research. However, there are few datasets that include the visual and tactile data. Sufficient visual, tactile, and position data can clearly describe the grasping process and improve the robot's ability of grasping.

According to the previous work, it is necessary to build a complete dataset for the robotic manipulation. In this chapter, a new grasp dataset based on the three-finger robot hand is built. In the following section, the structure of the multimodal dataset is introduced in detail. Moreover, the CNN and long short-term memory networks (LSTMs) are designed to complete grasp stability prediction.

2. Grasp stability prediction

In this section, the multimodal fusion framework of grasp stability prediction is proposed.

2.1 Visual representation learning

Under the visual image set, we can only observe 2700*2 = 5400 sets of image data in total, which is in use. It is difficult to extract visual features with convolutional neural networks (ResNet-18 network structure is used in our experiment). Training convergence is less on a small dataset, so time comparison network is used [10], capture video information from the capture process, anchor, positive, negative data. Then we define the triplet loss function [20] and use the characteristics of the continuous change of motion in the video to learn the operation process. The visual characteristics are also used as a pre-training process for the subsequent stable retrieval of the convolutional network part of the prediction network. Such as shown in **Figure 1**, we cleverly use a multi-angle camera to record the video image of the same capture process; at the same time, different image in the perspective should represent the same robot state, that is, its embedded layer embedding vector. A certain distance from the feature representation is relatively small, and the image at the same perspective at different times represents the robot. At different grasping states, a certain distance of the embedded layer Embedding vector is relatively large, formally:

$$\left\|f\left(x_i^a\right)-f\left(x_i^p\right)\right\|_2^2+\alpha<\left\|f\left(x_i^a\right)-f\left(x_i^n\right)\right\|_2^2 \tag{1}$$

where $f\left(x_i^a\right), f\left(x_i^p\right)$, and $f\left(x_i^n\right)$ represent the anchor, positive, and negative image features extracted by CNN. So, we can define the loss function [21] as

$$l(a,p,n)=\frac{1}{N}\left(\sum_{i=1}^{N}\max\left\{d\left(a_i,p_i\right)-d\left(a_i,n_i\right)+\alpha,0\right\}\right) \tag{2}$$

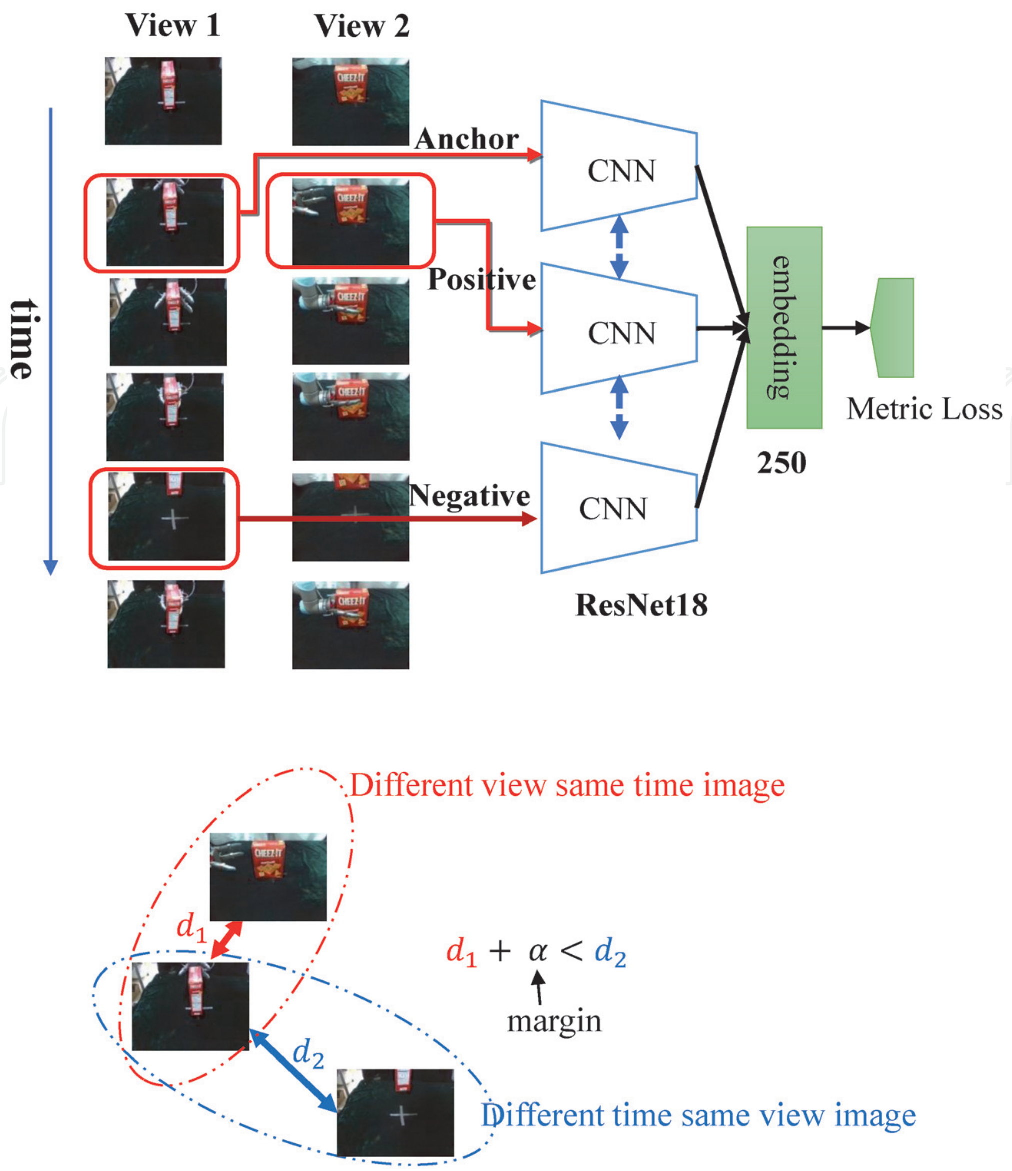

Figure 1.
Visual representation network.

2.2 Predicting grasp stability

In order to describe the properties of the objects like shape or size, the images are captured before grasping from two cameras, represented by Ib (**Figure 2**). Id is the position of the robot concerning the object grasped. Hence the vision feature fv can be calculated as

$$fv = R(Ib, Id) \tag{3}$$

where R represents the pre-trained neural network.

The images are passed through the standard convolutional network that uses the ResNet-18 architecture. Different from the previous work [22], the tactile sensors are used to obtain the force applied by the robot during the manipulation. As tactile sequences, the LSTMs are applied as the feature extractor:

$$ft = L(T0, T1, \dots, TT) \tag{4}$$

where ft is the last time step of the LSTMs' output and $T0, T1, \dots, TT$ is the input of the LSTMs at each step.

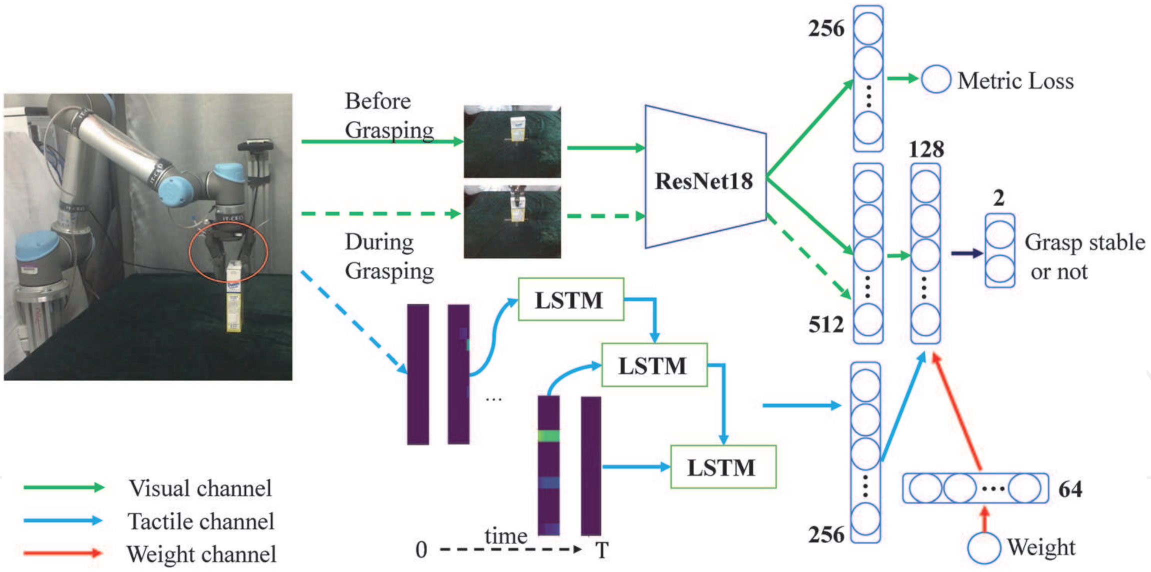

Figure 2.
Multimodal information predict grasp stability network.

Besides, the mass and mass distribution of the object also affect the stability of grasping. In order to simplify the problem, the weight of the object is known and the mass distribution is assumed uniform. Then the intrinsic object property is described as

$$fi = M(w) \tag{5}$$

where fi represents the intrinsic object feature and w is the object weight.

Then the multilayer perceptron (MLP) is used to extract the intrinsic feature. The sensory modalities provide the complementary information about the prospects for a successful grasp. For example, the camera's images show that the gripper is near the center of the object, and the tactile shows that the force is enough to keep stable for grasping. In order to study the method of multimodal fusion for predicting grasp outcomes, a neural network is trained to predict whether robot's grasp would be successful integrated by visual, tactile, and object's characters. The network computes $y = \{(X)\}$, where y is the probability of a successful grasp and $X = [fv, ft, fi]$ contains a set of images from multiple modalities: visual, tactile, and object intrinsic properties.

Train **the network**: initializing the weights of visual by CNN with a model pre-trained in Section III-A. The visual representation network is trained 200 epochs using the Adam optimizer [23], starting with a learning rate of 105 which is decreased by an order of magnitude halfway through the training process.

During the training, the RGB images are cropped with containing the table that holds the objects. Then, following the standard practice in object recognition, the images are resized to be 256 × 256 and randomly sampled at 224 × 224. Meanwhile the images are randomly flipped in the horizontal direction. However, the same data is still applied for augmentation to prevent overfitting.

3. Experiment and data collection

The experiment platform consists of the Eagle Shoal robot hand, two RealSense SR300 cameras, and the UR5 robot arm. As shown in **Figure 3**, they are arranged around the table of length 600 mm and width 600 mm. There is a layer of sponge on the surface of the table for protection. A soft flannel sheet covers the table to

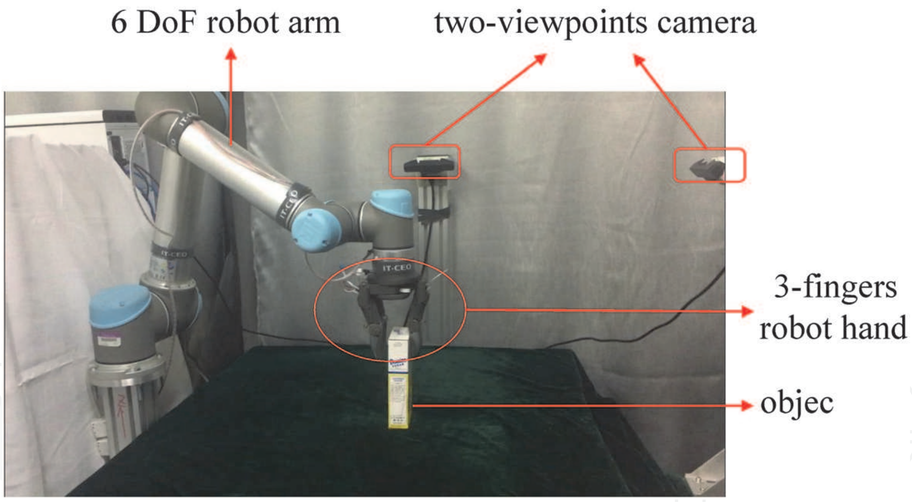

Figure 3.
Experiment platform.

Hand	Type	Weight	Force (mA)	Direction	Trail	Data type	Total
Eagle Shoal	10 objects	Empty	50/100/150	Top/right/back	10 times	T1/I	900 sets
Eagle Shoal	10 objects	Half/full	50/100/150	Top/right/back	10 times	T1/T2/I/V	1650 sets

Table 1.
Dataset statistics.

avoid the interference of light reflection. The UR5 robot arm with the Eagle Shoal robot hand fixes at the backside of the table. One RealSense camera is on the opposite side of the table for recording the front view of grasping. The other RealSense camera is located in the left of the table for recording the lateral view.

The general grasp dataset is built with various variables including shape, size, weight, grasp style, etc. The objects in the dataset contain different sizes of cuboid, cylinder, and special shapes, and their weights change by adding granules or water. Different grasping methods are tested by grasping from three directions including back, right, and top. The dataset with unstable grasping data is generated by slipping with added weight, changed grasp force, and adjusted grasp position (**Table 1**). The detailed processes are as follows:

1. The object is put on the center of the table; the front camera is used to get the point cloud data and computer the target's position.

2. Choose the object's half height position as the grasped point, control the robot to approach the object, and then add the random error of ±5 mm.

3. Based on the object's size, controlling the robot hand to grasp with a position loop mode, and then after 1 second, the robot arm lifts up with a speed of 20 mm/s.

4. After the robot arm moves to a certain position, the robotic finger's position is changed. If the hand is bending too much, this grasp is labeled as failure, and open the hand directly then prepare the next grasp.

5. The grasp is labeled as success, for a light object, if the robot puts down the object and, for some heavy object, the robot opens the hand and drops the object directly.

6. Putting the object on the center of the table, the robot arm returns to the initial place and waits for the next loop.

The proposed method is contrasted with traditional classifiers including k-nearest neighbor (KNN) [24], support-vector machine (SVM) [25], and naive Bayes (NB) [26]. A total of 2550 sets have been divided into 80% for training and 20% for testing. The KNN classifier is set with k = 3, and the SVM kernel is the radial basis function (RBF). The success rate with a criterion, the number of detection n and the number of label data m, is calculated by n/m. The contrast result in **Table 2** shows the performance of LSTM and SVM is both well with a success rate. However, the SVM's labels are on the falling edge, which means the SVM model gets a good classification result by learning the falling edge features. The falling edge means the object is dropped already and cannot help to realize a stable grasp. SVM proves unsuitable for this test.

Besides the success rate, another criterion is necessary to evaluate the slip detection. If the time of predict result turns from 1 to 0 ahead of the time in label data, set it as ahead sample and counted number n*ahead*, calculate the ahead rate by n*ahead*/m, and set it as the criterion. The results are shown in **Table 2**. With these two criteria, LSTM shows the superior performance that attains the higher success rate and higher ahead rate (**Figures 4** and **5**).

Classifier	Success rate	Ahead drop	Ahead forecast
KNN	0.7970	0.8176	0.4961
SVM	0.8467	0.6667	0.2569
NB	0.6881	0.6569	0.4843
OUR	0.9460	0.8588	0.6373

Table 2.
Classification results of different classifiers.

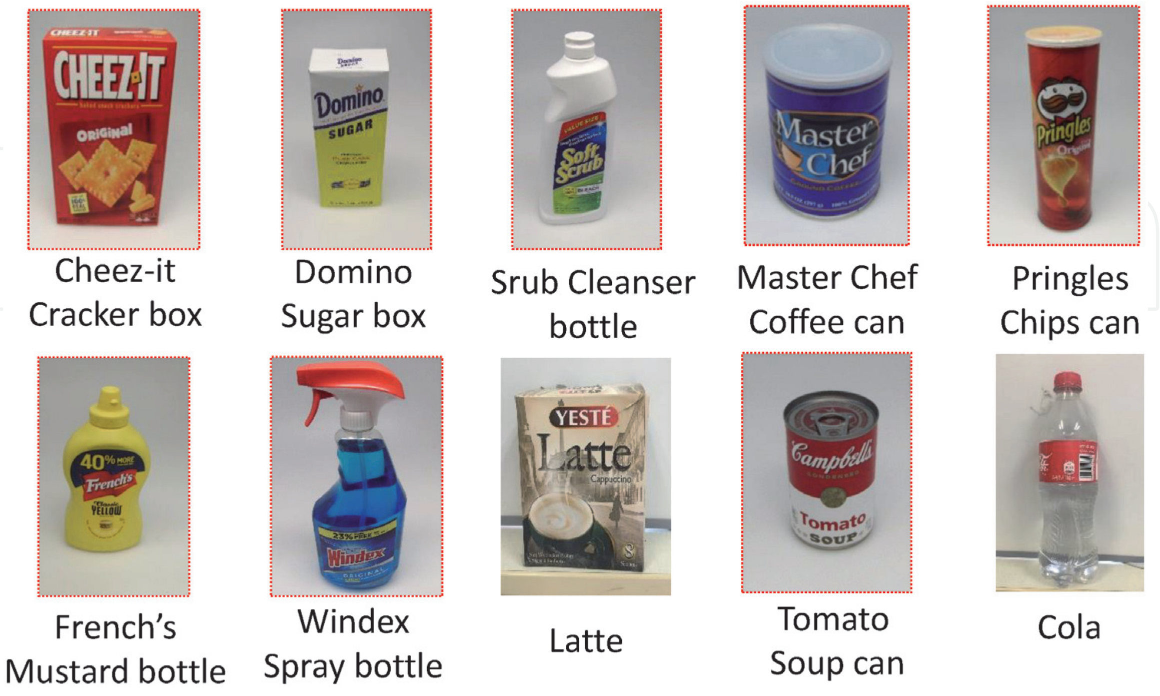

Figure 4.
All the grasp object, from YCB object set.

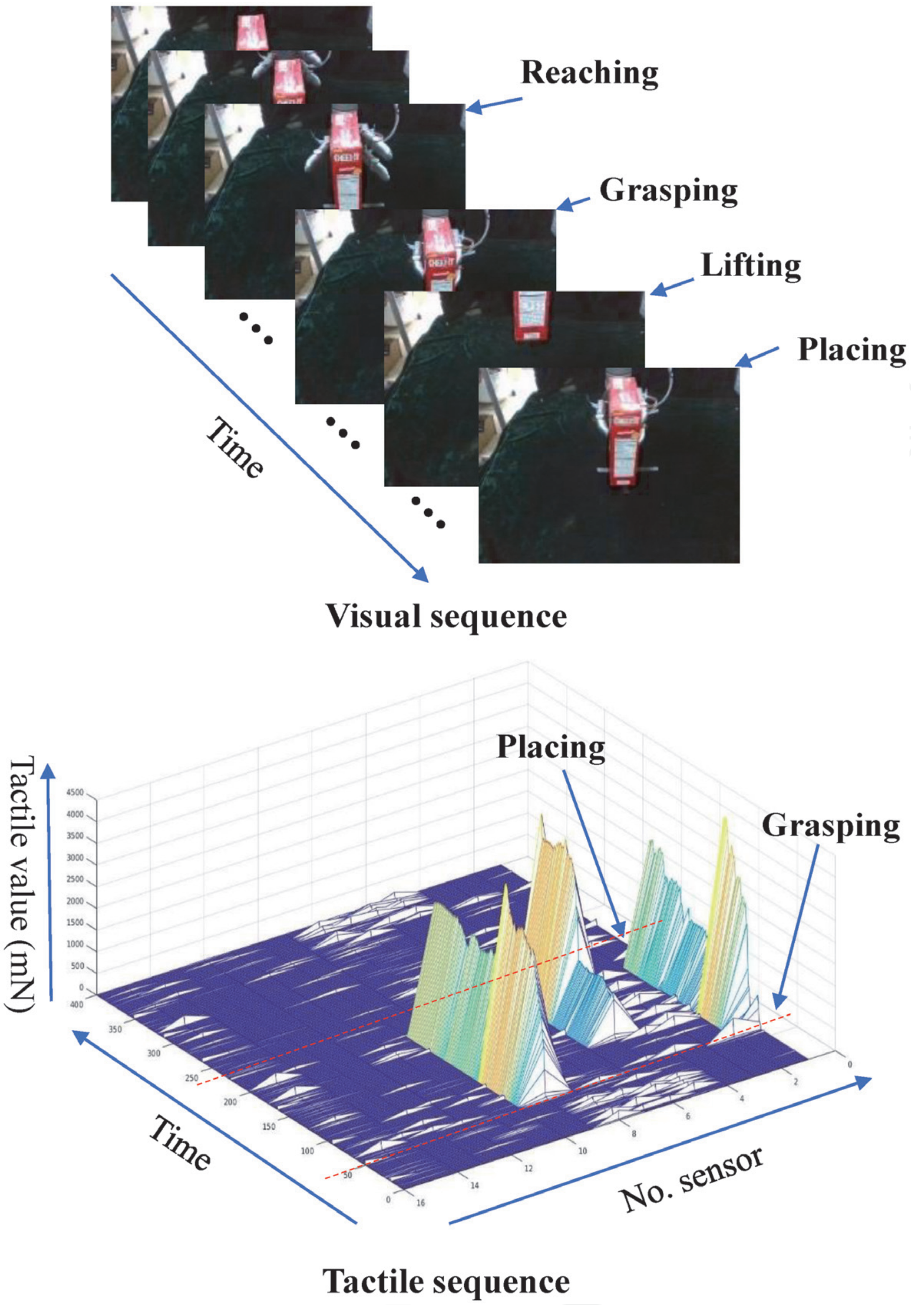

Figure 5.
Visual and tactile information visualization. Visual: grasping process video image sequence; and tactile: grasping process tactile sensor value.

4. Conclusions

In this chapter, the end-to-end approach for predicting stable grasp is proposed. Raw visual, tactile, and object intrinsic information are used, and the tactile sensor provides detailed information about contacts, forces, and compliance. More than 2500 grasp data are autonomously collected, and the multiple deep neural network model is proposed for predicting grasp stability with different modalities. The results show that visual-tactile fusion method improves the ability to predict grasp outcomes. In order to further validate the method, the real-world evaluations of the different models in the active grasp are implemented. Our experimental results demonstrate the superiority of the proposed method.

Acknowledgements

This work was supported by the Tsinghua University Initiative Scientific Research Program No. 2019Z08QCX15.

Author details

Bin Fang*, Chao Yang, Fuchun Sun and Huaping Liu
Department of Computer Science and Technology, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, China

*Address all correspondence to: fangbin@tsinghua.edu.cn

References

[1] Bohg J, Morales A, Asfour T, Kragic D. Data-driven grasp synthesis survey. IEEE Transactions on Robotics. 2014;**30**(2):289-309

[2] Levine S, Pastor P, Krizhevsky A, Ibarz J, Quillen D. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. The International Journal of Robotics Research. 2018;**37**(4–5):421-436

[3] Guo D, Sun F, Kong T, Liu H. Deep vision networks for real-time robotic grasp detection. International Journal of Advanced Robotic Systems. 2016; **14**(1):1-8

[4] Mahler J, Pokorny FT, Hou B, Roderick M, Laskey M, Aubry M, et al. Dex-Net 1.0: A cloud-based network of 3D objects for robust grasp planning using a multi-armed bandit model with correlated rewards. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2016. pp. 1957-1964

[5] Mahler J, Liang J, Niyaz S, Laskey M, Doan R, Liu X, et al. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. arXiv preprint arXiv: 1703.09312. 2017:1-12

[6] Mahler J, Matl M, Liu X, Li A, Gealy D, Goldberg K. Dexnet 3.0: Computing robust robot vacuum suction grasp targets in point clouds using a new analytic model and deep learning. arXiv preprint arXiv: 1709.06670. 2017:1-16

[7] Zhang T, McCarthy Z, Jowl O, Lee D, Chen X, Goldberg K, et al. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2018. pp. 1-8

[8] Vlack K, Mizota T, Kawakami N, Kamiyama K, Kajimoto H, Tachi S. Gelforce: A vision-based traction field computer interface. In: CHI’05 Extended Abstracts on Human Factors in Computing Systems. ACM; 2005. pp. 1154-1155

[9] Goldfeder C, Ciocarlie M, Dang H, Allen PK. The Columbia grasp database. In: IEEE International Conference on Robotics and Automation, 2009. ICRA’09. IEEE; 2009. pp. 1710-1716

[10] Sermanet P, Lynch C, Chebotar Y, Hsu J, Jang E, Schaal S, et al. Time-contrastive networks: Self-supervised learning from video. arXiv preprint arXiv:1704.06888. 2017:1-15

[11] Pinto L, Gupta A. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2016. pp. 3406-3413

[12] Li J, Dong S, Adelson E. Slip detection with combined tactile and visual information. arXiv preprint arXiv:1802.10153. 2018:1-6

[13] Kobayashi F, Kojima F, Nakamoto H, Kida Y, Imamura N, Shirasawa H. Slip detection with multi-axis force/torque sensor in universal robot hand. International Journal of Applied Electromagnetics and Mechanics. 2012;**39**(1–4):1047-1054

[14] Heyneman B, Cutkosky MR. Slip classification for dynamic tactile array sensors. The International Journal of Robotics Research. 2016;**35**(4):404-421

[15] Chebotar Y, Hausman K, Su Z, Molchanov A, Kroemer O, Sukhatme G, et al. BiGS: BioTac grasp stability dataset. In: ICRA 2016 Workshop on Grasping and Manipulation Datasets; 2016

[16] Chebotar Y, Hausman K, Su Z, Sukhatme GS, Schaal S. Self-supervised regrasping using spatio-temporal tactile features and reinforcement learning. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE; 2016. pp. 1960-1966

[17] Stachowsky M, Hummel T, Moussa M, Abdullah HA. A slip detection and correction strategy for precision robot grasping. IEEE/ASME Transactions on Mechatronics. 2016; **21**(5):2214-2226

[18] Kasper A, Xue Z, Dillmann R. The kit object models database: An object model database for object recognition, localization and manipulation in service robotics. The International Journal of Robotics Research. 2012;**31**(8):927-934

[19] Calli B, Walsman A, Singh A, Srinivasa S, Abbeel P. Benchmarking in manipulation research using the Yale-CMU-Berkeley object and model set. Robotics & Automation Magazine, IEEE. 2015;**22**(3):36-52

[20] Chechik G, Sharma V, Shalit U, Bengio S. Large scale online learning of image similarity through ranking. Journal of Machine Learning Research. 2010;**11**(Mar):1109-1135

[21] Chechik G, Sharma V, Shalit U, Bengio S. Large scale online learning of image similarity through ranking. Journal of Machine Learning Research. 2010;**11**:1109-1135

[22] Calandra R, Owens A, Upadhyaya M, Yuan W, Lin J, Adelson EH, et al. The feeling of success: Does touch sensing help predict grasp outcomes? arXiv preprint arXiv: 1710.05512. 2017:1-10

[23] Kingma DP, Ba J. Adam: A method for stochastic optimization. Journal of Computer Science. 2014:1-15

[24] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992;**46**(3): 175-185

[25] Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and Their applications. 1998;**13**(4):18-28

[26] John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc.; 1995. pp. 338-345