2017 IEEE International Conference on Robotics and Automation (ICRA)
Singapore, May 29 - June 3, 2017

A Hybrid Deep Architecture for Robotic Grasp Detection

Di Guo, Fuchun Sun, Huaping Liu, Tao Kong, Bin Fang and Ning Xi

***Abstract*— The robotic grasp detection is a great challenge in the area of robotics. Previous work mainly employs the visual approaches to solve this problem. In this paper, a hybrid deep architecture combining the visual and tactile sensing for robotic grasp detection is proposed. We have demonstrated that the visual sensing and tactile sensing are complementary to each other and important for the robotic grasping. A new THU grasp dataset has also been collected which contains the visual, tactile and grasp configuration information. The experiments conducted on a public grasp dataset and our collected dataset show that the performance of the proposed model is superior to state of the art methods. The results also indicate that the tactile data could help to enable the network to learn better visual features for the robotic grasp detection task.**

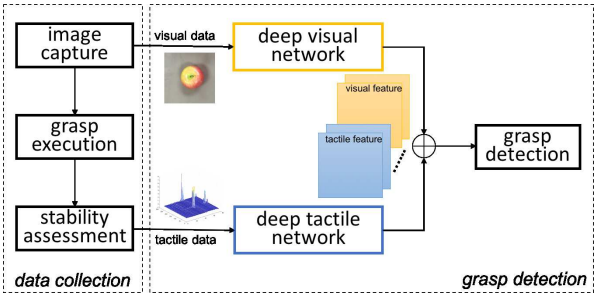

Fig. 1. A hybrid architecture for robotic grasp detection.

I. INTRODUCTION

With the development of sensor and machine learning techniques, robots are able to acquire more and more humanlike perception and abilities and thus to interact with the real world environment. As one of the most common and fundamental skills, robotic grasping has been widely studied during the past decades[1]. The goal of the grasp is to stably hold a target object and the whole process of the robotic grasping is composed of both the planning stage and the execution stage. For the grasp planning part, the mainstream approach is to find proper grasp configurations based on visual features from images [2]. However, some issues such as calibration problems, object attributes and hand restriction can influence the performance of the grasp. A seeming good grasp configuration predicted from visual information may still fail in the execution stage and vice versa. So it is necessary to investigate what other factors can help to generate a stable robotic grasp. It is noted that for each successful grasp, there is always physical contact between the object and the hand. So it is possible to resort to the tactile sensing to assess the stability of the real world robotic grasp and make a better grasp detection.

To find proper grasp locations for an object, numerous grasp planning approaches based on the 3D model [3][4] can be used to generate proper contact points and hand configurations. The possible solutions are then evaluated by certain quality metrics [5] to yield the optimal one and be executed. But it is usually very cumbersome to collect real world 3D data of the object. To overcome this problem, some synthetic object datasets are often used to test the grasp planning algorithms in a simulated environment [4][6][7]. However, in the real world, the model of the object is usually unknown and the reconstructed model is usually incomplete and inaccurate, which hinders the above approaches' application in the contact-level grasping in real scenario.

Considering these problems, it is necessary to develop a practical grasp planning method which can generate grasp configurations directly from the real world sensing data. Saxena et al. [2] proposed to learning grasping points directly from images, which doesn't require the reconstruction of the object's 3D model. It provides an alternative for the robot to implement the grasping with visual sensing information. Compared with the traditional range sensors which are responsible for the 3D reconstruction, the 2D image is far more easy to acquire and it doesn't suffers from problems when the object is reflective or transparent. Numerous work has put great effort on identifying proper grasp locations in the image. In [2][8], the local appearance features in the 2D images are used to find the graspable points on the object. Recently, witnessing the successful application of the deep vision networks in the area of computer vision[9][10], some work [11][12][13] tries to introduce the deep learning techniques to the area of robotic grasping. Having the images as the input, end-to-end frameworks are established to learn the grasping area from the images in real time. However, some important factors such as the objects attributes may be hidden from the visual information and the visual information may cheat in some cases. So knowing the extra information besides the visual information is important to generate more robust grasp strategies in the real world. Because of the physical contact between the hand and the object during the grasp, it is natural to employ the tactile sensing to facilitate the task of robotic grasping[14][15].

In this work, a hybrid deep architecture for real world robotic grasp detection is proposed (Fig. 1). The motivation for this architecture is to enhance the robot's perception

Di Guo, Fuchun Sun, Huaping Liu, Tao Kong and Bin Fang are with the Department of Computer Science and Technology, State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084, China. Di Guo and Ning Xi are also with the Department of Industrial and Manufacturing Systems Engineering, the University of Hong Kong, Hong Kong, China. Email: guodi.gd@gmail.com

ability and grasping skill in the real world environment. It is composed of the data collection phase and the grasp detection phase. During the data collection phase, both the visual and tactile data are collected. The tactile data is also used to assess the stability of the grasp which serves as the ground truth information for the grasp detection network. In the grasp detection phase, we introduce the tactile data into the visual network to strength the learning ability of the whole network. The contribution of this paper can be concluded as follows:

(1) A novel hybrid deep architecture is proposed for real world robotic grasp detection. It fuses the visual sensing and tactile sensing to generate better grasp detection result in the real world environment.
(2) A THU grasp dataset is collected for objects of different internal attributes and a robust method using tactile data is proposed to assess the stability of the grasp. The dataset contains the visual information, tactile information and grasp configurations.
(3) Experiments have been implemented both in a public dataset and our THU dataset, which illustrates the effectiveness of the proposed deep networks and architecture.

II. Related work

In the real word scenario ranging from the household environment to factory area, visual sensing is the most direct way for the robot to perceive its surroundings. So it is expected that the robot can know where to grasp the object by understanding the visual information. One of the important issues is to map the visual information into the robotic working space. Jiang et al. [16] propose to use a rectangle in the image to represent a parallel grasp, which unifies the grasp features in the image and the robotic hand configuration. Each of the two opposed edges of the rectangle denotes the parallel gripper's fingertip position.

It is difficult for the robot to determine some internal attributes of the object only with the visual information, while the internal attributes such as the material, stiffness, texture play important roles in robotic grasping. Object with different internal attributes may require different grasp configurations. Meanwhile, due to the physical contact between the robotic hand the object when a grasp is applied, a lot of work resorts to the tactile sensing to assess the stability of the robotic grasp. For the tactile sensing data collection, the experimental setup is usually fixed ignoring the uncertainty of the real world environment. In [17], the shape and the position of the objects is known beforehand. The object is classified as simple primitives and the grasp planner generates the specific pregrasp configurations. Numerous grasps are then applied to the object. Other approaches [18][19] hard-code the way the robotic hand grasping the known objects. The study of the robotic grasp stability is segmented from the robotic grasp planning phase. Recently, some remarkable work combine the grasp planning phase and the grasp assessment phase together to train deep networks to detect grasp configurations from images and execute the grasp [20], [21]. The self supervision approach is used to assess the success of the grasp. However, the grasp stability setup doesn't take the different properties of objects into consideration.

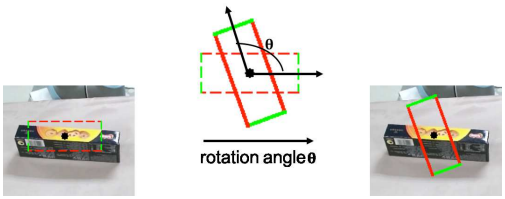

Fig. 2. A rotated grasp rectangle is used to represent the grasp configuration of a parallel grasp.

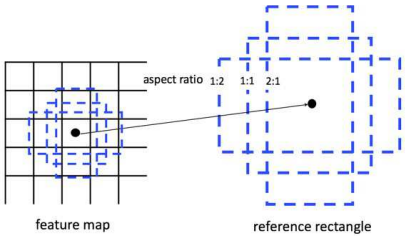

Fig. 3. The feature map and the reference rectangles. Reference rectangles of three ratios are used, which are denoted by dash lines. They sample each location across the feature map in a sliding window fashion.

III. Architecture

A novel hybrid deep architecture for real world robotic grasp detection (Fig. 1) is proposed. It combines the visual sensing and tactile sensing together to enhance the robot's perception ability and grasping skill.

It is composed of the data collection and grasp detection phase. In the data collection phase, the image of the scene is firstly captured and the object is extracted from the image. By knowing the position of the object, the robotic hand will approach to the object and apply the grasp in an arbitrary configuration. Meanwhile, the tactile sensing data is recorded and used to assess the stability of the grasp. During the data collection phase, both the visual data and tactile data are collected. It is worth noting that there is no human input in the whole process of the collection and the tactile sensing is used to assess each grasp instead of human labeling [16].

These mutimodal data collected can then be fed into the grasp detection phase, which contains a deep visual network and a deep tactile network. Because the tactile sensing can reveal some internal attributes of the object which are not perceptible from the visual sensing but important for robotic grasping. It is expected that the learning ability of the visual network will be strengthened by introducing the tactile sensing. Therefore, when a new object comes, although the tactile data is not available before grasping, the visual network will still give specific and accurate grasp detection result to guide the movement of the robot.

IV. Deep grasp detection networks

We use the grasp rectangle [16] to repesent the parallel grasp in an image. As shown in Fig. 2, the green edges denote the position of the opposed fingers. The black dot in the image denotes the grasp position of the object. To get to

the initial configuration, the robotic hand will approach to the grasp position vertically to the workspace and the initial grasp rectangle is represented in dash lines. Then the hand will rotate an angle of θ to adjust the pose of the object and grasp it. Therefore, the grasp position and the rotation angle can fully determine the grasp configuration.

A. Reference rectangle

Inspired by the Faster R-CNN [22] framework which integrates the region proposal network to the object detection network. We use the reference rectangle to efficiently indicate potential graspable candidates in the image. The reference rectangle samples all the possible locations in the convolutional feature map in a sliding window fashion and yields a grasp probability score for each rectangle. In Fig 3, reference rectangles with one scale and three aspect ratios are demonstrated. Each reference rectangle can then be refined to its corresponding grasp rectangle. The scale and aspect ratio of the reference rectangles may differ in different situations. This design of the reference rectangles enables the network to detect every possible location in the image and regress their corresponding grasp rectangles.

B. Model architecture

The architecture of our network is shown in Fig. 4. It is composed of three parts. The first part is responsible for feature extraction. We employ the Zeiler and Fergus (ZF) model [23] which has five shared convolutional layers to extract features from the input image. The ZF model shows great ability on object recognition task, which is believed to provide robust features from the image and proved to have the ability to generalize to different visual tasks. The last part is the grasp detection part which contains three sibling convolutional layers. It outputs the graspable score for each reference rectangle and its corresponding regressed initial grasp rectangle and the rotation angle for each initial grasp rectangle. For the output of the network, there are two class of labels for each reference rectangle, namely graspable and ungraspable. $\{t^x, t^y, t^w, t^h\}$ are the offset coordinates for the predicted initial grasp rectangle and $\{0^\circ, 10^\circ, \cdots, 170^\circ\}$ are the 18 labels for the rotation angle. The intermediate layer combines the feature extraction part and the grasp detection part. It is a $256-d$ feature map generated by a 3×3 kernel sliding across the previous layer. So we define the loss function for the reference rectangle as

$$L(g,a,t) = L_{grp}(g,\hat{g}) + \lambda_1 \hat{g} L_{ang}(a,\hat{a}) + \lambda_2 \hat{g} L_{reg}(t,\hat{t}) \quad (1)$$

where g is the graspability score for the reference rectangle and t is the offset parameters for the predicted initial grasp rectangle and a is the rotation angle class for the corresponding initial grasp rectangle. $\hat{g}$ is the ground truth label for the reference rectangle. It is 1 when the overlap between the reference rectangle and the ground truth grasp rectangle is above a specific threshold (0.5 in this work), otherwise it is valued as 0. And $\hat{t}$ is the offset parameters between the reference rectangle and the ground truth grasp rectangle and $\hat{a}$ is the ground truth rotation angle. The Softmax loss is used to evaluate if the reference rectangle is graspable or not and which angle class the rectangle belongs to. For the regression loss, we adopt the loss function (smooth L1) as described in [9] to regress the offset parameter t from the ground truth $\hat{t}$, which is proved to be easy to converge.

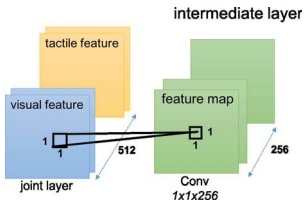

Fig. 5. The combination of the visual sensing and the tactile sensing.

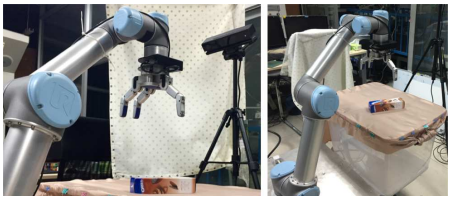

Fig. 6. The data collection platform.

For the tactile data, we utilize the data collected during the process of the grasp. The Barrett hand has three fingers and a 3×8 tactile array mounted on each fingertip. To better assess the quality of the grasp, we also use the strain reading from each finger. By concatenating the tactile sensing together with the strain reading, we get a 75-dimension tactile time sequence. For each time sequence may have different length, we employ the bilinear interpolation technique to resize the sequence into the same length. The tactile sensing data is then fed into a feature extraction convolutional network similar to the visual data. To combine the tactile sensing and visual sensing, we concatenate the feature map of each modal to obtain a joint layer and use a 1×1 kernel to sliding across the joint layer (Fig. 5). The resulted feature map can then be used as the intermediate layer in Fig. 4. Afterwards, the deep network can yield the grasp detection result.

V. Experiments

In order to verify the effectiveness of the proposed architecture. A THU grasp dataset is firstly collected. And then we apply the deep grasp detection network both on a public human labeled grasp dataset [16] and the collected dataset. We get the best grasp detection accuracy on the public dataset. Experimental results also demonstrate that the tactile sensing can help to learn better visual features so that to get better grasp detection result.

A. Data collection

The data collection setup is composed of a UR5 robotic arm, Barrett hand and the Kinect2. The Barrett hand has three identical fingers and we spread the two movable fingers to the position opposed to the fixed thumb. A modified storage box is used as a workspace. We use the cloth to cover an open storage box. The cloth stretches tightly and is fixed

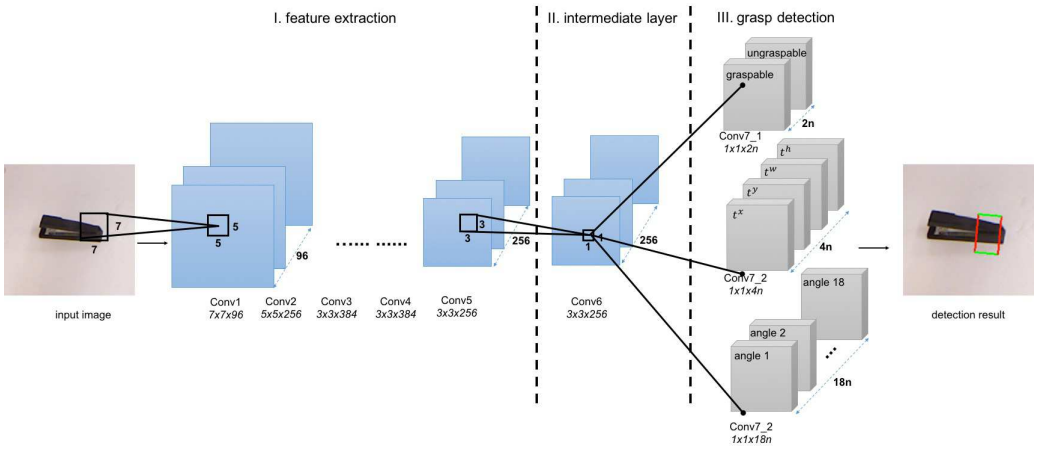

Fig. 4. The architecture of the deep visual network for the grasp detection. n is the number of the reference rectangle used in each location.

to the box edges by clips. This design is supposed to act as a cushion when the object drops and also protects the robotic finger when bad grasp planning occurs accidentally. The Kinect2 is placed in the front of the platform which can capture the visual information of the experimental scenario.

To better evaluate the stability of the grasp, we consider both the tactile and strain readings of the hand when it grasps the object. In most cases, a grasp is considered to be a stable one if the tactile reading from the fingertip is above a threshold value. However, for some soft objects such as dolls, the tactile reading is pretty small even when the hand has already successfully grasp the object. It is unfair to use only tactile reading as a unified criterion to judge each grasp. So we also take the strain value which is sensitive to the movement of the finger into consideration to assess the stability of the grasp. Therefore, we define the grasp assessment value q as:

$$q(t,s) = \lambda_t \sum_{i=1}^{n} \ln(t_i+1) + \lambda_s \sum_{i=1}^{n} \ln(s_i+1) \qquad (2)$$

where n is the number of fingers, t_i and s_i is the maximum tactile reading and strain gauge reading of the ith finger. λ_t and λ_s is scale factors for the tactile and strain sensing respectively, which is used to adjust the weight of each modality. They are defined as:

$$\lambda_t = \frac{1}{k_t \ln(t^*+1)}, \quad \lambda_s = \frac{1}{k_s \ln(s^*+1)}, \quad n(\frac{1}{k_t}+\frac{1}{k_s}) = 1$$

where t^* and s^* is the tactile and strain threshold value, k_t and k_s are the weight values.

The process of the data collection is shown in Fig. 7. An object is placed on the cloth surface and depth information is used to segmented the object from the background. After detecting the object, the arm moves vertically to the center of the object according to the depth information. The gripper rotates around the vertical axis of the palm with an arbitrary angle among $[0°, 10°, \cdots, 170°]$ and attempts to grasp the object. And then the fingers close gradually. Meanwhile, the tactile reading on the fingertip and the strain reading

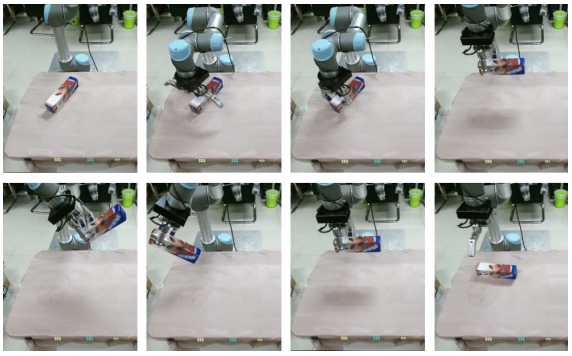

Fig. 7. The process of the data collection.

TABLE II

THE GRASP DETECTION RESULTS WITH DIFFERENT APPROACHES.

approach	detection accuracy (%)	
	instance-wise	object-wise
Jiang et al.[16]	60.5	58.3
lenz et al.[11]	73.9	75.6
Redmon et al.[12]	88.0	87.1
Ours	**93.2**	**89.1**

are recorded. The fingers stops moving when the grasp assessment value (q) is above a preset threshold value (0.6 in our experiment) or the fingers are fully closed. Then the hand will lift and swing right and left, which is to check if the object is stably grasped. The object will drop if the grasp is unstable. The grasp assessment value (q) will be calculated again after the checking process. The grasp will be labeled stable if the assessment value is still above the threshold value. Finally, the hand opens and the object drops on the cloth. The whole process is repeated continuously and autonomously without any human input unless the object's position is out of the camera's range. The visual data, tactile data and grasp configuration (grasp position and rotate angle) are all recorded. We choose 17 object with different material and each object is grasped by 100 times.

TABLE I

DIFFERENT REFERENCE RECTANGLE SETTINGS AND CORRESPONDING RESULTS OBTAINED IN THE EXPERIMENT.

#	setting	scale	aspect ratio	instance-wise				object-wise			
				20%	25%	30%	35%	20%	25%	30%	35%
a	1 scale; 1 aspect ratio	54^2	1:1	0.938	**0.932**	0.910	0.853	0.851	0.828	0.793	0.741
b	1 scale; 3 aspect ratio	54^2	1:2, 1:1, 2:1	0.932	0.921	0.893	0.859	0.879	0.851	0.822	0.759
c	3 scale; 3 aspect ratio	$27^2, 54^2, 108^2$	1:2, 1:1, 2:1	0.881	0.864	0.836	0.768	0.908	**0.891**	0.851	0.805

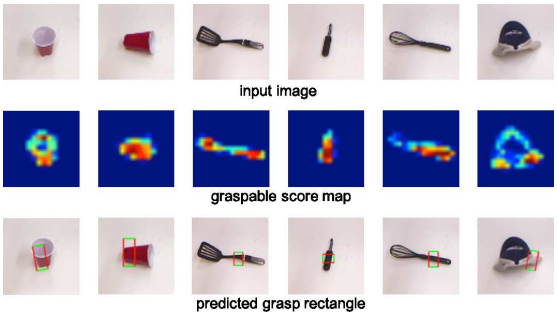

Fig. 8. Grasp detection results in the Cornell Grasp Dataset.

B. Grasp detection results and analysis

We firstly employ the deep visual model in Fig. 4 to analyze the detection results in both the public available Cornell Grasp Dataset and our THU dataset. In the Cornell Grasp Dataset, 280 household objects are placed arbitrarily on the tabletop and a total number of 885 images are collected. For each object in the image, several graspable rectangles and ungraspable rectangles are labeled by humans. The dataset is randomly split into the training set and testing set with a ratio of 4:1.

Given the image as the input, the model can yield the graspable score map and the predicted rectangle (Fig. 8). The first row of the picture demonstrates the input image. The second row is the corresponding graspable score map which denotes the grasp saliency of each object. The brighter (red) the color is, the more probable that location is suitable for the grasp. And the third row visualize the predicted grasp rectangle. We can see from the first two examples that the proposed model can learn different grasp configurations for the same object with different poses. Also, for objects having obvious handle parts, the predicted grasp rectangle falls at that very location.

Similar to [12], we employ the grasp IoU metric to evaluate the performance of the the grasp detection result. Assuming G is the detected grasp rectangle and G^* is the ground truth rectangle, then the grasp IoU can be defined as

$$IoU = \frac{G \cap G^*}{G \cup G^*} \tag{3}$$

where $G \cap G^*$ is the overlap area between the detected grasp rectangle and the ground truth rectangle and $G \cup G^*$ is the union area of these two rectangles. The predicted grasp rectangle is supposed to be a good one if the IoU value is above a threshold. We compare the performance of the proposed model with different reference rectangle settings with different IoU threshold values. The detailed results are listed in TABLE I and the curves are plotted as in Fig. 9 and Fig. 10. It can be concluded from the results that for the instance-wise case, the referenct rectangle setting with 1 scale and 1 aspect ratio has the best performance while for the object-wise case, reference rectangle setting with 3 scale and 3 aspect ratio performs best. The design of the reference rectangles should consider different application cases. When a known object appears, a simple reference rectangle setting is enough to get the ideal result. However, when an unknown object appears, it is better to have a thorough detection of the image with multiple reference rectangle settings.

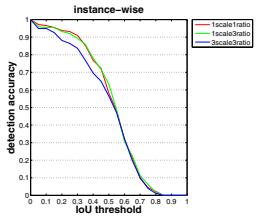

Fig. 9. The IoU threshold accuracy curve for the instance-wise case.

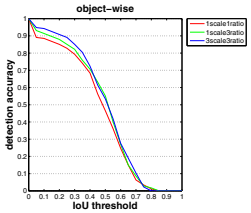

Fig. 10. The IoU threshold accuracy curve for the object-wise case.

Comparisons with different approaches on the same dataset have also been made. We consider the predicted grasp rectangle as a good one if the IoU between the predicted grasp rectangle and ground truth is above 25% and the rotation angle difference is within 30°. The comparison results are demonstrated in TABLE II. Our model gets the best performance among these approaches. Similar experiments have also been made in our THU dataset. The results are demonstrated in Fig. 11.

C. Multimodal grasp detection results

In order to verify that the tactile data can help to enhance the grasp detection performance. We compare the grasp detection result in THU grasp dataset. The 1 scale 1 ratio grasp reference rectangle setting performs best. The results are demonstrated in Fig. 12. It depicts the detection accuracy value along with the change of IoU threshold values for the visual only feature and both modalities. The multimodal model performs slightly better than the single modal model, which indicates that the tactile sensing is complementary to the visual modality and do help to strengthen the visual feature for grasp detection from the image.

However, we can see from Fig. 12 that the detection accuracy is lower than that is on the public dataset. It is because that for the public datset, multiple grasp rectangles are labeled for a single image, the IoU value is the largest

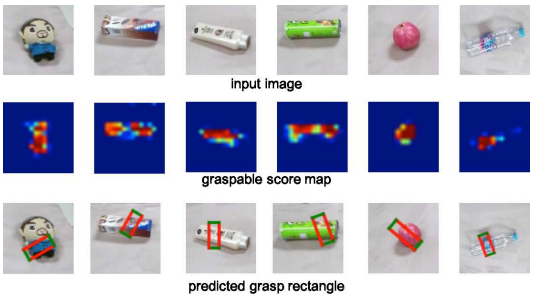

Fig. 11. Grasp detection results in THU grasp dataset.

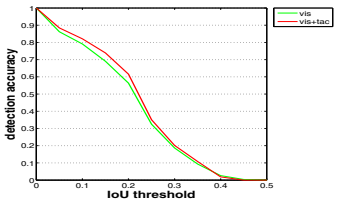

Fig. 12. The IoU threshold accuracy curve for both the visual feature and multimodal feature.

one among all the labeled grasp rectangles. While in the THU grasp dataset, we use the only one real robotic grasp configuration as the ground truth label, which makes the problem even more difficult.

VI. Conclusion

In this paper, we propose a novel hybrid deep architecture for detecting robotic grasps. It is robust to assess the grasp stability of object of different internal attributes. With this assessment method, a new THU grasp dataset is collected, in which the visual information, corresponding tactile data and grasp configurations are collected. The collection process is without any human input and completely from the perspective of the robot.

The deep network is also proposed to detect proper grasp rectangles from the visual and multimodal sensing information. The network has tested both on a public grasp dataset and our own dataset. It obtains the best results in the public dataset comparing with state of the art approaches. The experimental results demonstrate that the tactile data could help to enable the network to learn better visual features for the robotic grasp detection. In the future, the proposed model is also supposed to be used in a cluttered environment.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China under Grants 91420302, 61327809, 61210013, U1613212 and in part by the National High-Tech Research and Development Plan under Grant 2015AA042306.

References

[1] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-driven grasp synthesisa survey. *IEEE Transactions on Robotics*, 30(2):289–309, 2014.

[2] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, 2008.

[3] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *IEEE International Conference on Robotics and Automation*, pages 348–353. Citeseer, 2000.

[4] Corey Goldfeder, Matei Ciocarlie, Hao Dang, and Peter K Allen. The columbia grasp database. In *IEEE International Conference on Robotics and Automation*, pages 1710–1716. IEEE, 2009.

[5] Raúl Suárez, Jordi Cornella, and Máximo Roa Garzón. *Grasp quality measures*. Institut d'Organització i Control de Sistemes Industrials, 2006.

[6] Andrew T Miller, Steffen Knoop, Henrik I Christensen, and Peter K Allen. Automatic grasp planning using shape primitives. In *IEEE International Conference on Robotics and Automation*, volume 2, pages 1824–1829. IEEE, 2003.

[7] Sahar El-Khoury, Anis Sahbani, and Veronique Perdereau. Learning the natural grasping component of an unknown object. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2957–2962. IEEE, 2007.

[8] Quoc V Le, David Kamm, Arda F Kara, and Andrew Y Ng. Learning to grasp objects with multiple contact points. In *IEEE International Conference on Robotics and Automation*, pages 5062–5069. IEEE, 2010.

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[10] Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. Hypernet: towards accurate region proposal generation and joint object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 845–853, 2016.

[11] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[12] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In *IEEE International Conference on Robotics and Automation*, pages 1316–1322. IEEE, 2015.

[13] Di Guo, Tao Kong, Fuchun Sun, and Huaping Liu. Object discovery and grasp detection with a shared convolutional neural network. In *IEEE International Conference on Robotics and Automation*, pages 2038–2043. IEEE, 2016.

[14] Huaping Liu, Yunhui Liu, and Fuchun Sun. Robust exemplar extraction using structured sparse coding. *IEEE transactions on neural networks and learning systems*, 26(8):1816–1821, 2015.

[15] Huaping Liu, Di Guo, and Fuchun Sun. Object recognition using tactile measurements: Kernel sparse coding methods. *IEEE Transactions on Instrumentation and Measurement*, 65(3):656–665, 2016.

[16] Yun Jiang, Stephen Moseson, and Ashutosh Saxena. Efficient grasping from rgbd images: Learning using a new rectangle representation. In *IEEE International Conference on Robotics and Automation*, pages 3304–3311. IEEE, 2011.

[17] Yasemin Bekiroglu, Janne Laaksonen, Jimmy Alison Jorgensen, Ville Kyrki, and Danica Kragic. Assessing grasp stability based on learning and haptic data. *IEEE Transactions on Robotics*, 27(3):616–629, 2011.

[18] Hao Dang and Peter K Allen. Learning grasp stability. In *IEEE International Conference on Robotics and Automation*, pages 2392–2397. IEEE, 2012.

[19] Qian Wan, Ryan P Adams, and Robert D Howe. Variability and predictability in tactile sensing during grasping. In *IEEE International Conference on Robotics and Automation*. IEEE, 2016.

[20] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *IEEE International Conference on Robotics and Automation*. IEEE, 2016.

[21] Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *arXiv preprint arXiv:1603.02199*, 2016.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[23] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014*, pages 818–833. Springer, 2014.