An architecture for grasping complex-shaped, thin and reflective objects

Olivia Nocentini
Excellence in Robotics & AI
Scuola Superiore Sant'Anna
Pisa, Italy
olivia.nocentini@santannapisa.it

Jaeseok Kim
Excellence in Robotics & AI
Scuola Superiore Sant'Anna
Pisa, Italy
jaeseok.kim@santannapisa.it

Zain Muhammad Bashir
Excellence in Robotics & AI
Scuola Superiore Sant'Anna
Pisa, Italy
muhammadzain.bashir@santannapisa.it

Filippo Cavallo
Industrial Engineering
Università di Firenze
Firenze, Italy
filippo.cavallo@unifi.it

***Abstract*—In this paper, we proposed an architecture that integrates the Generative Grasping Convolutional Neural Network (GG-CNN2) with a few-shot segmentation model for identifying a suitable grasp pose. The proposed architecture is validated by performing real-world grasping and pick and place experiments. Our framework achieved a success rate of over 85.6% on the pick and place task of seen surgical tools and 90% on unseen surgical tools.**

***Index Terms*—Generative Grasping Convolutional Neural Network; Few-shot segmentation; Surgical tools**

I. Related Work

Recently, deep learning methods [1] have proved effective in robotic grasp generation [2]. In order to accelerate the detection speed, the recent GG-CNN network proposed by Morrison et al. [3] directly evaluated the grasp quality and pose of grasps at every pixel. The same authors improved the GG-CNN creating a new network for real-time grasp prediction called GG-CNN2. This architecture is a fully convolutional network based on the semantic segmentation architecture from Yu and Koltun [4], which uses dilated convolutional layers to provide improved performance in semantic segmentation tasks. Semantic segmentation is one of the most common computer vision tasks, which segments the specific object from the input image with classification. Li et al. proposed a new network architecture, which is few-shot segmentation model (FS-Seg), integrated VGG-16 as a backbone and relation module that produces a segmentation map [5]. The network input is the support set (image samples with annotations) and query set (image samples without annotations). The output is the per-pixel classification and segmentation of the query image. The few-shot segmentation method is useful in our work because we have few surgical tools. Therefore, it is applied for the surgical tool's segmentation from RGB images that can calculate the depth (z value) during the grasping process.

Concretely, the contributions of our work are the following:

- Creating a new dataset of surgical tools composed of only complex-shaped, thin and reflective objects that are usually less easy to grasp due to difficult depth estimation resulting from their thinness;
- Proposing an architecture for grasping complex-shaped, thin and reflective objects as surgical tools using the GG-CNN2 with a segmentation method that provides the depth of surgical tools images;
- Validating the architecture for grasping the surgical tools of seen and unseen in non cluttered or cluttered environments.

Identify applicable funding agency here. If none, delete this.

II. Experimental Setup

Our experimental set-up uses a Universal Robot (UR5) equipped with a Robotic gripper, and a Kinect v2 is used to acquire the images of the surgical tools. Since our dataset size is very small compared to datasets such as the Cornell dataset and the Jacquard dataset, we augmented it using random rotations, zooming, jitter and white noises, and brightness variations. At the end, we obtained a dataset of 9920 RGB images (the original dataset consists of 320 RGB images). The augmented dataset will be publicly available.[1].

We used seen (used for training) and unseen objects (27 and 9 respectively) to evaluate quantitatively the grasping of the surgical tools using the GG-CNN2 with or without the segmentation method. We used a 80/20% split for training/testing of the GG-CNN2.

The color segmentation method using OpenCV is used to collect a dataset for FS-Seg. Based on the original RGB data from GG-CNN2, we prepared 224 x 224 RGB and a binary image of each tool and made at least five examples that have a different position and orientation of each tool (Total of 320 RGB and binary images). We employed a pre-trained model, which already extracted features from the FSS-1000 dataset [5] that improve the segmentation with our small dataset.

During the experimental phase, we tested the performance in grasping not only of the GG-CNN2 but also of the GG-CNN to compare the two networks. From the training of the GG-CCN and GG-CNN2, we obtained several models, based on different combination of augmentation data. We used those models to predict grasp poses of each surgical tool with the GG-CNN and GG-CNN2 with or without segmentation. For each configuration, we grasped each surgical tool 10 times.

[1]The dataset will be released on github: https://github.com/OliviaNocentini/SurgicalKit

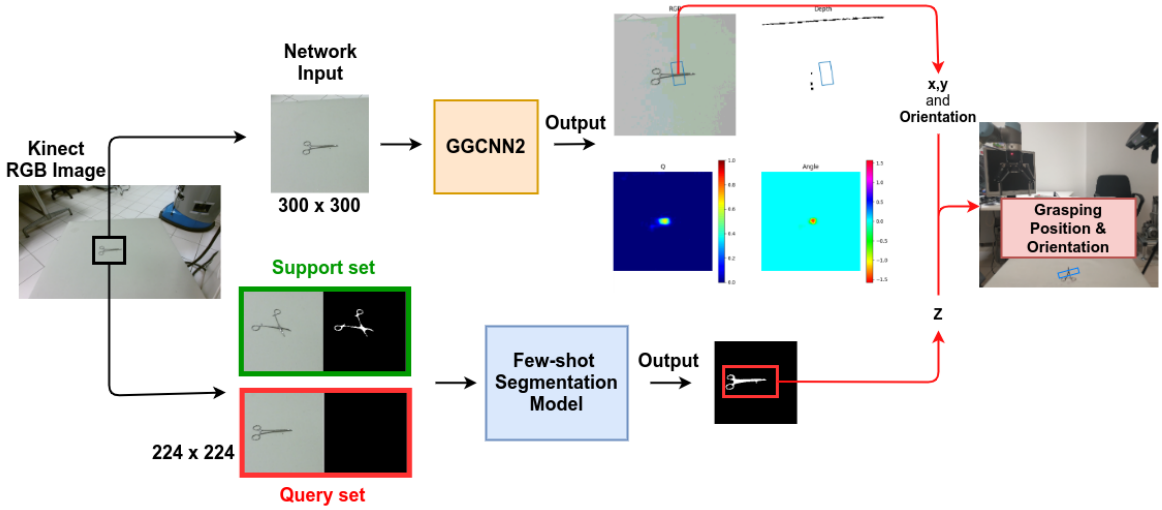

Fig. 1. Overview of our proposed architecture. From the Kinect RGB image, 300 × 300 image as input to the GG-CNN2 network for the tool detection and grasping's position (x and y) and orientating calculation. Then, 224 × 244 image of support and query set as input to the few-shot segmentation model for calculating average depth z value based on the segmentation.

III. Experimental Results and Discussions

We evaluated Intersection over Union (IoU) with GG-CNN/GG-CNN2 during training. The best IOU of each network is 71% and 76% respectively. We also evaluated IOU of FS-Seg model and it is 59.15%. The IOU of FS-Seg model is not a high value, but the model could predict the segmentation of input query image in real-time environment.
Regarding the GG-CNN without the segmentation method, we obtained a success rate of 62% in grasping seen and (73.64%) in grasping unseen surgical tools. The unexpected fact with the GG-CNN is that the grasping with unseen objects has a better performance compared to the one with the seen objects. As regards the GG-CNN2, a success rate of 94.8% and 78.18% respectively in grasping seen and unseen tools is observed. The GG-CNN2 without segmentation has a better performance both on seen and unseen objects compared to the GG-CNN and this is due to its dilated convolutional layers which provide improved performance in semantic segmentation tasks.

As concerns the GG-CNN and the segmentation method, a success rate of 74% for the seen and 66.36% for the unseen is obtained. For what regards the GG-CNN2 with the segmentation method, the performance obtained was of the 85.6% for seen and the 90% for unseen objects. Therefore, we concluded that the segmentation method together with the grasping for unseen surgical tools has better results than without the segmentation.

We also evaluated the grasping performance of multiple surgical tools. We used the unseen surgical tools (9 items) and we tested 9 times of one configurations of the unseen objects.

During the experimental part with a single surgical tool, we faced the following common problems: the most of the issues were related to the falling off of the tool after being grasped. This happened due to the large length and no consideration of center of mass of the objects. Also it is occurred due to the higher weight of the surgical tools and the material they are made of. The surgical tools' shape is also very important for the success or failure of grasping since strange shapes are more difficult to be grasped and thick objects are easier to be taken compared to the thickness ones. Other two failure cases are that GG-CNN2 cannot find grasping box from input data with different colored tools, and also FS-Seg model gives a wrong segmentation. Both of failure cases occurred due to the lack of examples during the training.

For what concerns the issues found in grasping multiple surgical tools, we faced the following problems. First, sometimes the grasping box is wrong located and we couldn't grasp any surgical tool. Secondly, due to the material or weight of the surgical tool, there is the falling off of it after its grasping. Finally, there is the wrong computation of the surgical tool depth because we used the average of depth of the output from FS-Seg model. As a consequence, it could be the failure in grasping the object.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] 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.

[3] D. Morrison, P. Corke, and J. Leitner, "Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach," in *Proc. of Robotics: Science and Systems (RSS)*, 2018.

[4] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[5] X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, and C.-K. Tang, "Fss-1000: A 1000-class dataset for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2869–2878.