Generative Perception for Robotic Grasping

Douglas Morrison[1,2], Peter Corke[1,2], and Jürgen Leitner[1,2].

I. Introduction

The ability for a robot to grasp and manipulate previously unseen objects is an essential skill to be able to perform meaningful tasks in unstructured environments. While the problem of grasp synthesis has been investigated for many decades, the biggest advancements have been seen recently with the proliferation of vision-based deep-learning techniques – with a focus on anti-podal grasping – and a number of associated grasping datasets [1], [2], [3], [4].

Many of these techniques share a similar pipeline, where grasp candidates are produced by sampling an input image at discrete intervals of offset and rotation and ranked individually by a neural network classifier [1], [2], [4]. Alternatively, a single neural network pass can be used to generate grasp rankings at predetermined offsets and rotations [3] or generate a single pose per image [5] without the need to sample sections of the image. In each case, the output of such a system is the best detected grasp pose, which can then be executed by a robot. Broadly speaking, these techniques treat the problem of grasp synthesis as mainly one of detection and remain independent from the act of execution on the robot.

In contrast, we envision a generative model which directly encodes a one-to-one mapping from the image space to the grasping space. We present here our system design (Fig. 1) and preliminary results. Our system does away with the concept of grasp candidate sampling by directly generating an anti-podal grasp pose and success certainty for every pixel in an input image. When combined with depth information, the model outputs can be projected into 3D space, allowing a robot to perceive its environment in terms of grasp affordances.

Our generative approach improves on existing methods by better representing a robot's environment with respect to the distribution of grasp affordances than methods which rely on sampling the grasping space or generate only a single grasp pose per view. Additionally, this representation allows us to leverage other robotics mainstays such as visual servoing and active perception during reaching towards a target object, as opposed to reaching without feedback after an initial grasp pose estimation.

Furthermore, we intend to extend this concept to multi-fingered grasping. As such, we have collected a video dataset of human grasping on a small set of items.

II. Model for grasp point generation

Our generative model, shown in Fig. 1, is based on a convolutional autoencoder topology. Due to the large number of datasets available, we have chosen to first test our system on anti-podal grasps. Our model receives a 4-dimensional RGB-D image and produces 3 simultaneous outputs: a heatmap representing the likelihood of success for a grasp executed at every pixel and the rotation of the ideal grasp at each position (θ), encoded as the x ($cos\theta$) and y ($sin\theta$) components of a unit vector to avoid discontinuities in the output which could hinder training.

The output of the model can be transformed into a set of 6-DoF grasp poses by using depth information to produce 3D translations for every pixel in the image and aligning the grasp approach to the estimated surface normal, a common technique used by other grasping systems [1], [4].

We trained our model using data from the Cornell Grasping Dataset [6], which provides a set of human-labelled grasp poses for a diverse set of items. We created a set of training images for our model consisting of binary masks representing the centres of the positive labelled grasps from the dataset, and images encoding the angular components at the same location. 20% of the dataset images were reserved for testing.

In order to compare our grasp point detection performance against existing work, we produce a set of grasps for the set of test images by considering local maxima in the grasp score output, shown in Fig. 2 alongside the ground truth labels. The de facto standard for evaluation against the Cornell Grasping Dataset is to consider a success if the predicted grasping rectangle aligns within 30° of and has an intersection-over-union (IoU) of greater than 25% with a ground truth grasp. By this metric, our system has an accuracy of 82.8% when considering the global maximum over the grasp score output and 87.7% when considering the top two local maxima. For comparison, the current state of the art for this dataset is 89.21% [5].

However, we note that this metric may not be the best qualifier of system success. Because the Cornell Grasping Dataset is human labelled, it contains only a small subset of all possible grasps. As can be seen in the last row of Fig. 2, it is possible to generate seemingly valid grasp poses which don't correspond to any ground truth grasp labels (the tin-opener handle, the wrapped portion of the power brick and the camera cord). We look forward to validating our system on a physical robot.

III. Extending to multi-fingered grasping

We believe that our system can be adapted to generate grasp poses for multi-fingered robotic hands by generating

This research was supported by the Australian Research Council Centre of Excellence for Robotic Vision (ACRV) (project number CE140100016).

Contact: douglas.morrison@hdr.qut.edu.au

[1] Authors are with the Australian Centre for Robotic Vision (ACRV).

[2] Authors are with the Queensland University of Technology (QUT).