Robotics and Autonomous Systems 100 (2018) 194–205

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Refining object proposals using structured edge and superpixel contrast in robotic grasping

Lu Chen, Panfeng Huang *, Zhou Zhao

National Key Laboratory of Aerospace Flight Dynamics, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, PR China
Research Center for Intelligent Robotics, School of Astronautics, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, PR China

HIGHLIGHTS

- Eliminate the common assumptions made in grasping scenario, such as the fixed object position and monotonous manipulation background.
- A translation-to-alignment mechanism to align input proposals.
- Introduce the distance bias to improve the connectivity between superpixels.
- Experiments are performed on both benchmark dataset and robotic task.

ARTICLE INFO

Article history:
Received 22 September 2017
Received in revised form 26 October 2017
Accepted 27 November 2017
Available online 5 December 2017

Keywords:
Robotic grasping
Object detection
Structured edge
Superpixel
Local cues

ABSTRACT

Grasp detection is an active research branch in robotic field. Most existing works have made strong assumptions, such as the fixed object position and monotonous manipulation background, which facilitate the detection of graspable objects. But the real manipulation condition could be much more complicated. In this work, we propose a novel object perception method. It is able to accurately detect the object, as well as those in cluttered background, and guide the movement of robotic arm to reach a proper grasping state. First, we translate and align the initial proposals according to the structured edge distribution. The aligned proposals have a larger overlap with ground truth at the expense of a little drop in precision. Then, for each superpixel inside the proposal, we use its contrast to high-contrast superpixels and background superpixels, weighted by distance bias, to determine whether it should be included in the refined proposal. Experimental results on both benchmark dataset and robotic task have verified the effectiveness of the proposed method.

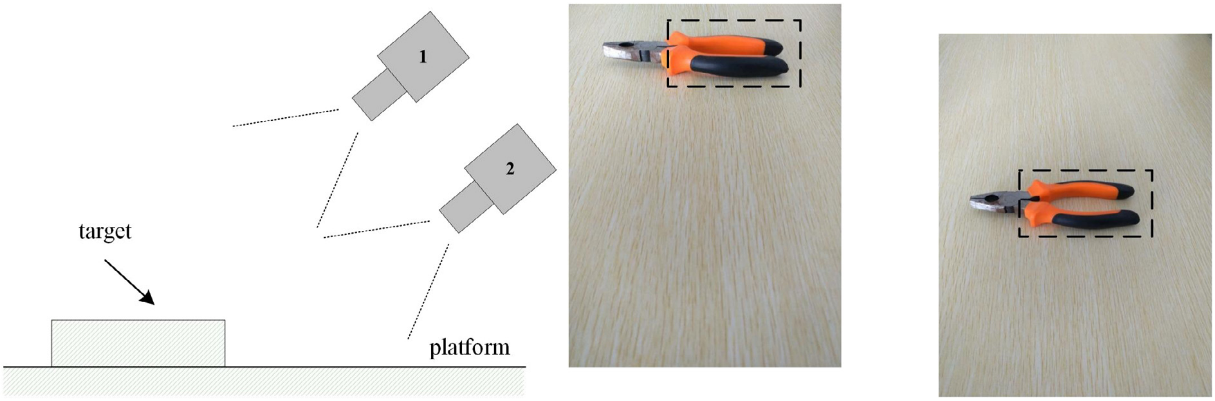
1. Introduction

With recent advances in robotic intelligence, the robots are capable of perceiving the environment better and have been applied into wider applications, such as household service, human–machine interaction, industrial manufacturing and on-orbit service [1]. In recent years, robotic grasping has become the fundamental problem and received numerous research concerns [2–4], for it builds the interaction between robot and its surroundings. A typical grasping procedure generally include three successive modules, object detection, grasp detection and gripper execution. In the first module, the object is extracted from the whole image. In the second module, we predict the object grasps, followed by driving the robotic arm and accomplishing the grasp task in the third module.

Considering the robotic arm control methods are quite mature, the accurate detection of object grasps has become the key technique. Recent studies generally treat it as a detection problem and predict the grasps directly from visual features [5]. [6] introduced a new dataset with large volume (∼50 k data points) and used it to train a Convolutional Neural Network for grasp detection. In addition to the visual features, tactile data can also be taken into consideration. In [7], measurements from tactile sensors served as time sequences to represent the dissimilarity between good and bad grasps. [8] adopted the tactile data to adjust the configuration of fingers to achieve a more stable grasp. In [9], both visual and tactile data were integrated to fuse information of multiple modalities. The experiments showed that this strategy was beneficial for object recognition. However, the extracted features could be less informative and difficult to distinguish good grasps if the object is imaged with limited view, as shown in (a) of Fig. 1, where no representative features could be used. When we move the camera

* Corresponding author at: Research Center for Intelligent Robotics, School of Astronautics, Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, PR China.
E-mail addresses: chenlu11@mail.nwpu.edu.cn (L. Chen), pfhuang@nwpu.edu.cn (P. Huang), zhaozhou@mail.nwpu.edu.cn (Z. Zhao).

https://doi.org/10.1016/j.robot.2017.11.015
0921-8890/© 2017 Elsevier B.V. All rights reserved.



(a) View 2.

(b) View 1.

Fig. 1. The imaging relationships between viewing camera and the target before and after the movement of robotic arm. (a) and (b) are images captured under these two viewing relationships. The object parts marked by the black dashed rectangles are identical but appear with distinct visual feature.

using the robotic arm and make sure the target locates in image center, a much better imaging view is achieved and the structure of the object appears. It contributes to the selection of a more robust and stable grasping point.

To the best of our knowledge, most existing methods do not take the influence of imaging view into consideration. Therefore, in most experimental sets of robotic grasping, the object is usually placed near the center of camera view [10,11], which is a high location prior. Another typical experimental set is that the object is usually placed on a monotonous background and could be easily located by subtracting the background. But in practice, the manipulation condition could be more complex and the objects' positions could be more random. If the object is poorly localized or missed, the robotic grasping task tends to fail. Hence, this work mainly focuses on the first module of the grasping procedure. Inspired by this, we propose an object detection method that is insensitive to cluttered background and object position. Once the object is localized, the robotic arm will be moved to the position which is vertically on the object.

Determining the position of all the objects in image, known as the object detection, is a classic research field in computer vision. Solutions to this problem contribute to the development of image understanding and artificial intelligence, e.g., surveillance [12,13], pedestrian detection [14,15], robotics [16] and space manipulation [17]. A large set of approaches have been proposed and conventionally, the object detection methods follow the sliding window paradigm, which suffers the heavy computational burden in exhaustive search, around 10^6 classifier evaluations per image [18]. In order to alleviate the computational complexity, the object proposals have been used to facilitate the judgment of objects. Compared with randomly selected windows, object proposals have a higher probability of covering objects in image. For an input image, the number of object proposals can range from $\sim 10^2$ to $\sim 10^4$ depending on which object proposal generation method it chooses.

As defined in [19,20], objects are standalone things with well-defined boundary and distinct from amorphous background stuff. According to this assumption, the similar visual properties shared by objects contribute to distinguish them from their surroundings. Note that object proposals are usually based on low-level features, including shapes, colors, textures, edges, etc., which enable the proposal generation methods rapidly exclude a large amount of background windows. By introducing the cascade structure, the speed and accuracy of evaluating proposals can be improved [21].

One problem cause by bounding box metric is that the proposals generated by most existing object proposal generation methods (e.g., [22], BING [23], Edge Boxes [19], Objectness [24], etc.) are difficult to coincide well with ground truth, as shown in Fig. 2. As the IoU threshold increases, the detection rate will decrease rapidly.

In order to rectify these position biases, a large amount of researches have been proposed in recent years. They can be roughly divided into two categories: regression-based methods and refinement-based methods. [25] proposed a new loss function and regressed four boundaries of the predicted bounding boxes simultaneously. [26] used every pixel of the feature map to regress a 4-dimensional vector as the border of the prediction. But regression-based methods need to be integrated into the training framework and iteratively improve the detection accuracy during training procedure. Refinement-based methods use bounding boxes produced by object proposal methods as the input and refine them only relying on local features. Hence, they can be directly concatenated into the available object proposal frameworks and improve the detection rate. [27] introduced the straddling degree to evaluate superpixels overlapped by the proposal borders. If the majority of a superpixel locates inside the proposal, the corresponding border will be adjusted to make sure the proposal can cover the whole superpixel. In [28], the bounding boxes were refined from the aspect of saliency detection. The candidate boxes were selected by ranking the saliency scores and the refined boxes were achieved by non-maximum suppression. The adobe boxes in [29] utilized color histogram as the contrast cue and determined whether each superpixel belongs to the object via local-contrast analysis. [30] introduced the depth information from RGBD image to boost the confidence in superpixel determination. [31] combined the LBP (Local Binary Pattern) feature and enclosed contour distribution together to produce more accurate object proposals.

In this paper, we first improve the quality of object proposals from the aspect of computer vision. The improved proposals could localize the objects with higher accuracy. Then, the proposed method is applied in the scenario of robotic grasping. Our work is similar to [29] but makes contributions in the following two aspects.

1. We introduce the translation-to-alignment mechanism before object refinement. In [29], the predicted proposals are directly used without any processing procedure. Considering the unavoidable position biases of proposals, we introduce the structured edge map and use it to align the input proposal first. Experiments show that it could achieve a larger overlap with ground truth.

2. We introduce the distance bias to improve the connectivity between superpixels. Considering only color histograms were considered in [29], object parts which have similar color distributions