Vision-Guided Tactile Poking for Transparent Object Grasping

Jiaqi Jiang[1], Guanqun Cao[2], Aaron Butterworth[2], Thanh-Toan Do[3], and Shan Luo[1]

I. INTRODUCTION

Transparent objects are widely used in our daily life, e.g., glass cups, plastic bottles and glass pan lids in a kitchen. They are also common in research laboratories, e.g., vials and glass flasks. However, it is still challenging for a robot to detect and grasp transparent objects. Compared to opaque objects, transparent objects lack salient features in their surfaces such as colour and texture features. Moreover, transparent materials do not adhere to the geometric light path assumptions made in classic stereo vision algorithms. This results in the depth information of transparent objects from depth sensors such as Intel RealSense D415 and D435 being inaccurate or with unpredictable noise. Due to these challenges, most of the current grasping methods that rely on accurate depth information from cameras cannot be directly applied to the grasping of transparent objects.

Humans grasp objects with rich sensory information, such as the visual information obtained from eyes and the tactile feeling via physical interaction. It is common that vision with a wide field of view is used first for fast localisation of objects, then touch providing accurate perception of compliance and contact force is used to align hand posture or grip strength to enable a stable grasp.

Inspired by those observations, we propose a novel vision-guided tactile poking approach for grasping transparent objects in this paper, as shown in Fig. 1. We first train a deep neural network named *PokePreNet* with synthetic RGB images to predict the *poking regions* that are with similar surface normals to the table surface. The contacts with those areas contribute to good tactile readings while leading to minimal disturbance to the state of the object. A robotic arm equipped with a tactile sensor is then guided to contact those regions, so as to generate informative local profiles of the contacted transparent objects. Finally, using the improved profiles, a heuristic grasp proposal is generated for grasping the transparent object.

II. METHODOLOGY

A. Poking Region Segmentation.

The poking region segmentation is treated as an instance segmentation problem. In the instance segmentation, every pixel will be simultaneously classified whether it belongs to the poking region and which instance it is part of. One of the most popular instance segmentation techniques is Mask R-CNN. However, the poking region only occupies a small part of the bounding box, which causes a bad precision of Mask R-CNN. To solve this issue, our PokePreNet introduces two novel improvements to the original Mask R-CNN for segmenting the poking regions: (1) a larger output feature map via adding more deconvolutional layers; (2) a new pixel-level Positive-Negative-balanced loss.

In the vanilla Mask R-CNN, the average binary cross-entropy loss is used for training instance masks. However, the distribution of positive/negative pixels (positive pixels are the pixels that are part of the poking regions, and negative pixels are ones that are not) in the objects is heavily biased: only 5% of the bounding box area is part of the poking region. To this end, the cross-entropy loss from the poking region only contributes to a small part of the total loss and leads to a bad precision of the poking region.

To address this issue, we define the following pixel-level Positive-Negative-balanced (PN) loss for the poking region mask L_{mask}:

$$\begin{aligned} L_{mask}(X_i) = -\beta_i \sum_{j \in Y_i^+} \log \Pr\left(y_j = 1 | \, X_i\right) \\ - \sum_{j \in Y_i^-} \log \Pr\left(y_j = 0\right) | X_i) \end{aligned} \tag{1}$$

where Y_i^+ and Y_i^- denote the positive and negative ground truth label sets for the i^{th} RoI X_i, respectively; β_i is the weight on an instance basis to balance the loss between positive and negative pixels. Specifically, β_i is set to $|Y_i^-|/|Y_i^+|$ and 1 when $|Y_i^+|$ is larger than 0 and equal to 0, respectively. $|.|$ function is used for calculating the set size, and j represents the pixel index. $\Pr\left(y_j = 1 | \, X_i\right) = \sigma\left(a_j\right) \in [0, 1]$ is computed using sigmoid function $\sigma\left(.\right)$ on the activation value a_j at pixel j.

B. Vision-guided Tactile Poking.

Given the detected poking region, we generate a poking point $\boldsymbol{P_t} = [x_t, y_t]$ in the image frame for every transparent object. To generate the poking point, we first find the external contour of the poking region mask using OpenCV function *findContours*. Then we use OpenCV function *fitEllipse* to fit the contour and get the centroid $\boldsymbol{P_c}$. As shown in the output of our PokePreNet in Fig. 1, if the poking region is a simply connected mask, the poking point will be set to the ellipse centroid $\boldsymbol{P_c}$, as centroids are widely used for grasping the objects with simple rectangular or cylindrical shapes. On the other hand, if the poking region is a ring shape mask, $\boldsymbol{P_c}$'s nearest positive pixel will be set as the poking point to avoid getting the GelSight sensor [1] into the object.

[1]J. Jiang and S. Luo are with Department of Engineering, King's College London, London WC2R 2LS, United Kingdom. E-mail: {jiaqi.1.jiang, shan.luo}@kcl.ac.uk. [2] G. Cao and A. Butterworth are with the smARTLab, Department of Computer Science, University of Liverpool, Liverpool L69 3BX, United Kingdom. [3]T.-T. Do is with Department of Data Science and AI, Monash University, Clayton, VIC 3800, Australia.

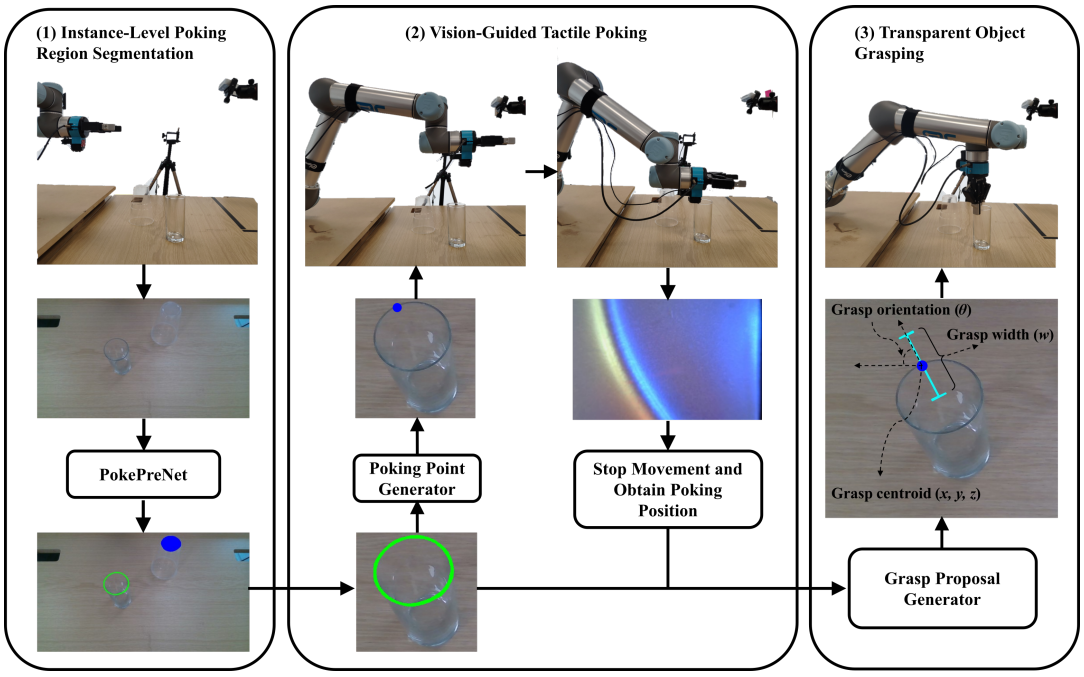

Fig. 1: An overview of our vision-guided tactile poking approach for transparent object grasping. **From left to right:** First, the PokePreNet takes the RGB image and outputs the segmented poking regions where different colours represent different instances. Then based on the detected poking regions, the poking point generator is used to generate the potential poking point that guides the robotic arm to move towards the transparent object until the equipped GelSight sensor contacts the object. Lastly, with predicted poking region and the obtained local profiles from tactile poking, a heuristic grasp proposal is generated for grasping the transparent object.

C. Heuristic Transparent Object Grasping.

Based on the predicted poking region and the object's local profiles (i.e., contact position) from the tactile poking, a heuristic grasp representation in the world frame is generated for the top-down parallel grasping. The grasp representation is defined as a 5-dimensional vector $\boldsymbol{G_{hrst}} = [x, y, z, w, \theta]$ as shown in Fig. 1, where $[x, y, z]$ represents the grasp centroid in the world frame. w and θ represent the width and the orientation of the heuristic grasp, respectively.

If $\boldsymbol{P_c}$ belongs to the poking region, the poking position $\boldsymbol{P_t^W}$ in the world frame will be equal to the position of centre $\boldsymbol{P_c^W}$. Hence, a centroid-based grasp is used for grasping the transparent object. If $\boldsymbol{P_c}$ is not part of the poking region, the grasp centroid will be set according to the distance $D(\boldsymbol{P_c^W}, \boldsymbol{P_t^W})$ between $\boldsymbol{P_c}$ and $\boldsymbol{P_t}$ in the world frame. If D is larger than half of the finger width, the gripper finger could be inserted into the transparent object. Hence, an edge grasp is used for grasping the transparent object as the grasp proposal shown in Fig. 1. Otherwise, a centroid-based grasp is used and the grasp position will be set to $\boldsymbol{P_c^W}$.

III. Experiment Results

To evaluate the performance of our PokePreNet, we construct a high-quality synthetic dataset as well as a real-world test benchmark with over 9,000 RGB images and their corresponding ground truth annotations. To bridge the gap between the simulation and the real world, we randomise the simulator to expose the model to a wide range of environments while training.

We evaluate PokePreNet on both synthetic and real-world benchmarks. Our experiments demonstrate that our proposed method can learn vision-guided tactile poking regions, with a high mean Average Precision (mAP) of 0.360, which outperforms 0.319 mAP achieved by the standard cross-entropy loss approach. Moreover, our method demonstrates the ability to generalise to transparent objects in real-world scenarios with solely synthetic data used for training.

We also conduct real robot experiments and the results show that the poking region is a better cue for guiding the tactile poking compared to the bounding box and instance mask. The better poking region segmentation contributed by our PN loss can further improve the poking success rate from 84.3% to 89.8%. Moreover, our proposed method can enhance the success rate of transparent object grasping from 38.9% to 85.2%, compared to a vision-based grasping method. Thanks to its simplicity, the proposed method can be adapted to other settings that use other force or tactile sensors such as GelTip [2] and Tactip [3], and can also be used for grasping other challenging objects. In future work, we will investigate the tactile alignment for grasping transparent objects without prior knowledge of the object shape.

References

[1] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[2] D. F. Gomes, Z. Lin, and S. Luo, "Geltip: A finger-shaped optical tactile sensor for robotic manipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9903–9909.

[3] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.