Visuotactile In-Hand Pose Estimation

Felix Nonnengießer[1], Alap Kshirsagar[2], Boris Belousov[3,4] and Jan Peters[2,3,4,5,6]

***Abstract*— This paper presents an approach to robotic in-hand object pose estimation, combining visual and tactile information to accurately determine the position and orientation of objects grasped by a robotic hand. We address the challenge of visual occlusion by fusing visual information from a wrist-mounted RGB-D camera with tactile information from vision-based tactile sensors mounted on the fingertips of a robotic gripper. Our approach employs a weighting and sensor fusion module to combine point clouds from heterogeneous sensor types and control each modality's contribution to the pose estimation process. We use an augmented Iterative Closest Point (ICP) algorithm adapted for weighted point clouds to estimate the 6D object pose. Our experiments show that incorporating tactile information significantly improves pose estimation accuracy, particularly when occlusion is high. Our method achieves an average pose estimation error of 7.5 mm and 16.7 degrees, outperforming vision-only baselines by up to 20%. To validate the practical applicability of our method, we conducted an insertion task experiment, demonstrating the ability to perform precise object manipulation in a real-world scenario.**

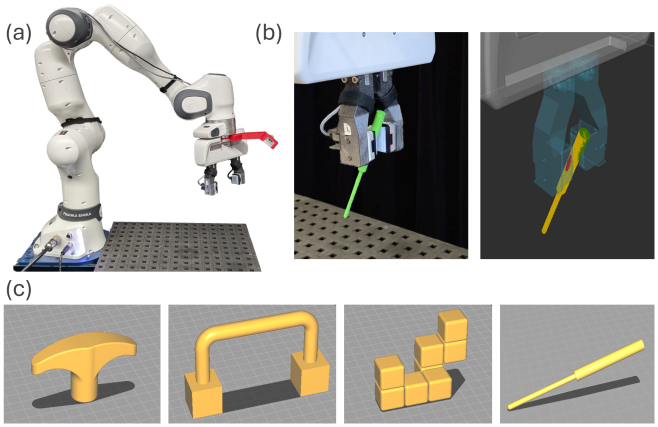

Fig. 1. **(a)**: We used the Franka Research 3 robotic arm with a parallel gripper equipped with GelSight Mini sensors and a RealSense D405 depth camera. **(b)**: A tool grasped by the gripper and its estimated pose along with the point clouds obtained by the sensors. **(c)**: Objects used in our experiments: *Knob, Handle, SL-Block, Screwdriver*

I. INTRODUCTION

In-hand pose estimation describes the process of determining the position and orientation of an object held within a robotic hand. This capability is crucial for robotic object manipulation and assembly tasks. To address this challenge, researchers have explored various approaches using visual and tactile information. Vision-based methods typically employ RGB or RGB-D cameras and use techniques such as feature and template matching [1], [2], point cloud registration [3], and machine learning approaches [4], [5]. However, these methods often struggle with occlusion caused by the grasping hand. Tactile-based approaches leverage different types of tactile sensors to infer object poses. Recent developments in camera-based high-resolution tactile sensing technologies like GelSight [6] and DIGIT [7] have enabled new methods in this domain [8]–[10].

Visuotactile-based approaches combine the strengths of both visual and tactile methods to enhance pose estimation accuracy and robustness [11], [12]. These approaches face the challenge of fusing dissimilar sensor modalities, often employing techniques such as extended Kalman filters [13] or neural networks [11], [14]. Some methods obtain an initial pose estimation based on visual data and then refine it using tactile information [15], [16].

This work presents a novel approach to robotic in-hand object pose estimation that integrates both visual and tactile information. Unlike many current pose estimation approaches that rely on deep learning techniques or require extensive training and large datasets, our method focuses on one-shot pose estimation without relying on learning-based algorithms or initial pose estimates. By utilizing a registration algorithm, we aim to create a flexible and adaptable approach for in-hand pose estimation that can be applied to new objects without extensive training or data collection.

II. PROPOSED METHOD

Our proposed method for robotic in-hand pose estimation combines visual and tactile sensor data to achieve accurate and robust results. We preprocess point clouds obtained by a depth camera and vision-based tactile sensors, fuse and apply different weights to the point clouds from different sources, and use an augmented Iterative Closest Point (ICP) algorithm adapted to handle weighted point clouds.

A. Data Processing

We use an Intel RealSense D405 Depth Camera attached to the end effector and directed towards the gripper (Fig. 1a). The camera provides an RGB color image and a depth stream, which is converted into a point cloud. This point cloud is segmented and filtered in subsequent steps, to obtain the points belonging to the object.

[1]Department of Computer Science, Goethe Universität Frankfurt, Germany. felixnon@cs.uni-frankfurt.de

[2]Intelligent Autonomous Systems Lab, Department of Computer Science, TU Darmstadt, Germany

[3]German Research Center for AI (DFKI)

[4]Department of Psychology, University of Giessen, Germany

[5]Centre for Cognitive Science, TU Darmstadt

[6]Hessian Center for Artificial Intelligence (Hessian.AI), Darmstadt

We thank Hessisches Ministerium für Wissenschaft & Kunst for the DFKI grant and "The Adaptive Mind" grant.

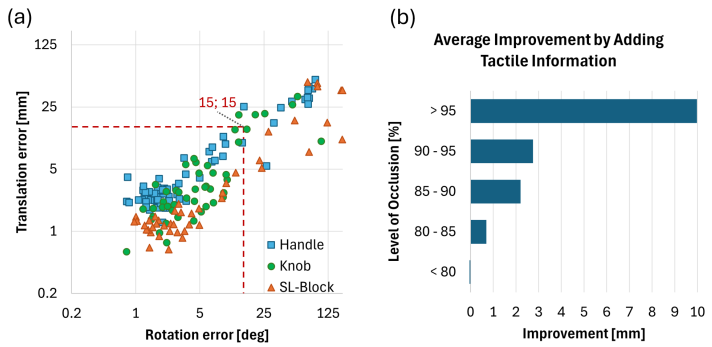

Fig. 2. **(a)**: Rotation and translation error of each pose estimation attempt using the object specific weight. **(b)**: The average improvement achieved by combining tactile and visual information compared to using visual information alone.

TABLE I

RESULTS FROM THE INSERTION TASK EXPERIMENT FOR DIFFERENT TOOL ANGLES AND RESULTING OCCLUSION. COMPARING PERFORMANCE OF USING VISION DATA ONLY (VIS) AND VISUOTACTILE INFORMATION (VIS+TAC).

Tool Angle	Vis		Vis+Tac		Occ.
0°	✓	✓	✓	✓	80%
-5°	✓	✓	✓	✓	82%
-10°	✓	✓	✓	✓	85%
-15°	✓	✓	○	✓	86%
-20°	○	○	✓	✓	86%
-25°	○	✓	✓	✓	88%
-30°	○	○	✓	○	89%
-35°	○	×	✓	○	93%
-40°	×	×	×	✓	98%
-45°	○	×	✓	✓	98%
-50°	×	×	✓	✓	98%

✓ Perfect insertion | ○ Successful insertion | × Failed insertion

Both fingers of the gripper are equipped with a GelSight Mini sensor, which provides details about the contact and local geometry of the grasped object. We reconstruct the deformation of the sensors' gel surface and generate a point cloud representing the local geometry of the grasped object.

B. Sensor Fusion and Point Cloud Weighting

We obtain two point clouds, one from the tactile module and one from the camera module. Based on the sensor modality, we assign a label and weight to each point to determine the influence each sensor modality should have during the subsequent pose estimation process. We fuse the sensor data by concatenating the individual point clouds into a new single point cloud while keeping track of each point's corresponding sensor modality and assigned weight.

C. Point Cloud Registration

We address the problem of object pose estimation with a point cloud registration algorithm to align the mesh of the original object with the point clouds obtained from the employed sensors (Fig. 1b). We augmented the conventional Iterative Closest Point (ICP) [17], [18] algorithm to allow for aligning weighted point clouds, accommodating multiple sensor modalities, and thus determining their influence during the alignment process. Since the ICP algorithm's success is dependent on the quality of the initial alignment, the registration is performed multiple times with varying initialization transformations.

III. EXPERIMENTS AND RESULTS

We conducted three experiments to evaluate our method. We tested on different objects with differing geometric features and shape (Fig. 1c). We compared the performance when using visual information only with the performance when employing a camera and tactile sencors combined.

(1) Weighting Strategy: We investigated the optimal weighting ratio between visual and tactile sensor modalities to achieve the highest accuracy in pose estimation. Results showed that the pose estimation outcome is influenced by the weighting of the sensor modalities and that the optimal weights varied for different objects. For the objects used in this experiment, we determined the following optimal vision-to-tactile weight ratios: *Knob:* 1:12.5; *Handle:* 1:3.5; *SL-Block:* 1:0.5. Further, we investigated dynamic weighting strategies to adapt weight ratios based on point cloud metrics such as occlusion and noise. While we were not able to formulate a universal dynamic weighting strategy, the results indicate a strong potential for improvement through such an adaptive system.

(2) Object Pose Estimation: We evaluated the accuracy and reliability of our method, investigated the benefit of combining visual and tactile information, and explored the impact of visual occlusion (Fig. 2). Incorporating tactile data reduced the average translation error from 9.44 mm to 7.50 mm, and rotation error improved from 21.09° to 16.70°. We observed a strong positive correlation between pose estimation error and visual occlusion of the object ($r(157) = .590$, $p < .001$). Especially in cases of high occlusion the additional tactile data provided significant improvement compared to the vision-only condition.

(3) Insertion Task: We performed a practical insertion task to demonstrate the real-world applicability of our method. A screwdriver was grasped at varying angles and inserted into a hole under different occlusion conditions. With low occlusion levels, visual information alone was sufficient to successfully insert the tool. However, in cases of high visual occlusion, the additional tactile information was necessary to compensate for missing visual features to allow successful insertions of the screwdriver (see Table 1).

IV. CONCLUSION

Our research has shown that combining visual and tactile information can significantly improve pose estimation accuracy and robustness compared to vision-only approaches. The addition of the tactile modality is particularly effective in scenarios with high visual occlusion. Future work could focus on developing an adaptive weighting strategy and testing the method on a wider variety of objects.

REFERENCES

[1] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "Deepim: Deep iterative matching for 6d pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 683–698.

[2] S. Zakharov, I. Shugurov, and S. Ilic, "Dpod: 6d pose object detector and refiner," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1941–1950.

[3] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[4] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[5] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6d object pose prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 292–301.

[6] 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.

[7] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[8] R. Li, R. Platt, W. Yuan, A. Ten Pas, N. Roscup, M. A. Srinivasan, and E. Adelson, "Localization and manipulation of small parts using gelsight tactile sensing," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 3988–3993.

[9] M. Bauza, A. Bronars, and A. Rodriguez, "Tac2pose: Tactile object pose estimation from the first touch," *The International Journal of Robotics Research*, vol. 42, no. 13, pp. 1185–1209, 2023.

[10] G. M. Caddeo, N. A. Piga, F. Bottarel, and L. Natale, "Collision-aware in-hand 6d object pose estimation using multiple vision-based tactile sensors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 719–725.

[11] T. Anzai and K. Takahashi, "Deep gated multi-modal learning: In-hand object pose changes estimation using tactile and image data," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9361–9368.

[12] Y. Gao, S. Matsuoka, W. Wan, T. Kiyokawa, K. Koyama, and K. Harada, "In-hand pose estimation using hand-mounted rgb cameras and visuotactile sensors," *IEEE Access*, vol. 11, pp. 17 218–17 232, 2023.

[13] G. Izatt, G. Mirano, E. Adelson, and R. Tedrake, "Tracking objects with point clouds from vision and touch," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4000–4007.

[14] S. Dikhale, K. Patel, D. Dhingra, I. Naramura, A. Hayashi, S. Iba, and N. Jamali, "Visuotactile 6d pose estimation of an in-hand object using vision and tactile sensor data," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2148–2155, 2022.

[15] Y. Liu, X. Xu, W. Chen, H. Yuan, H. Wang, J. Xu, R. Chen, and L. Yi, "Enhancing generalizable 6d pose tracking of an in-hand object with tactile sensing," *IEEE Robotics and Automation Letters*, 2023.

[16] A. N. Chaudhury, T. Man, W. Yuan, and C. G. Atkeson, "Using collocated vision and tactile sensors for visual servoing and localization," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3427–3434, 2022.

[17] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[18] Y. Chen and G. Medioni, "Object modelling by registration of multiple range images," *Image and vision computing*, vol. 10, no. 3, pp. 145–155, 1992.