NormalFlow: Fast, Robust, and Accurate Contact-based Object 6DoF Pose Tracking with Vision-based Tactile Sensors

Hung-Jui Huang[1], Michael Kaess[1], and Wenzhen Yuan[2]

***Abstract*— Tactile sensing is crucial for robots aiming to achieve human-level dexterity. Among tactile-dependent skills, tactile-based object tracking serves as the cornerstone for many tasks, including manipulation, in-hand manipulation, and 3D reconstruction. In this work, we introduce NormalFlow, a fast, robust, and real-time tactile-based 6DoF tracking algorithm. Leveraging the precise surface normal estimation of vision-based tactile sensors, NormalFlow determines object movements by minimizing discrepancies between the tactile-derived surface normals. Our results show that NormalFlow consistently outperforms competitive baselines and can track low-texture objects like flat table surfaces and ping pong balls. Additionally, we present state-of-the-art tactile-based 3D reconstruction results, showcasing the high accuracy of NormalFlow. We believe NormalFlow unlocks new possibilities for high-precision perception and manipulation tasks that involve interacting with objects using hands.**

I. Introduction

The skill to interact with and manipulate objects often relies on accurate in-hand object tracking capability, which remains a challenge for vision-based systems due to occlusions during manipulation. Vision-based tactile sensors like GelSight [1] offer a promising alternative, enabling occlusion-free tracking. Prior works [2], [3], [4], [5] handle object tracking by converting tactile images into point clouds and applying registration methods like ICP [6], but these often perform poorly due to the noise and distortion in tactile-derived point clouds. In this work, we introduce NormalFlow, a state-of-the-art tactile tracking algorithm that outperforms point cloud registration methods in both accuracy and speed. By directly minimizing discrepancies between surface normal maps—rather than relying on point clouds—NormalFlow achieves fast, robust, and accurate 6DoF pose estimation without object models, even on low-texture surfaces like flat table surfaces and ping pong balls. It achieves a mean translation error of 0.29mm (over a total movement of 3.4mm), rotation error of 1.9° (over a total movement of 37.4°), running at 70Hz on CPU. We also demonstrate its application in tactile-based 3D reconstruction, producing high-quality geometry. We believe NormalFlow opens new avenues for higher precision tactile-dependent perception and control.

II. Method

NormalFlow tracks object motion by directly minimizing differences between surface normal maps extracted from tactile images. We adapt the approach from [7] to estimate these maps. Let $\mathbf{I}$ and $\mathbf{I}'$ denote the surface normal maps of a reference and a target sensor frame, respectively, where each map is a function $\mathbb{R}^2 \mapsto \mathbb{R}^3$ from pixel coordinates to surface normals (Fig. 2). Our goal is to estimate the 6DoF transformation from the reference frame to the target frame $(\mathbf{R}_{\boldsymbol{\theta}}, \mathbf{t}_{\boldsymbol{\theta}}) \in SE(3)$, parameterized as $\boldsymbol{\theta} = (x, y, z, \theta_x, \theta_y, \theta_z) \in \mathbb{R}^6$. NormalFlow minimizes the difference between the transformed reference map $\mathbf{I}$ and the target map $\mathbf{I}'$ within the shared contact region:

$$\sum_{(u,v)\in\overline{C}} [\mathbf{I}'(\mathbf{W}(u, v; \boldsymbol{\theta})) - \mathbf{R}_{\boldsymbol{\theta}}\mathbf{I}(u, v)]^2 \tag{1}$$

where (u, v) is the pixel coordinates and $\overline{C}$ is the shared contact region. The re-mapping function $\mathbf{W}(u, v; \boldsymbol{\theta})$ maps pixel coordinates from the reference frame to the target frame. Inspired by the Lucas-Kanade optical flow method [8] [9], NormalFlow employs the Gauss-Newton optimization to

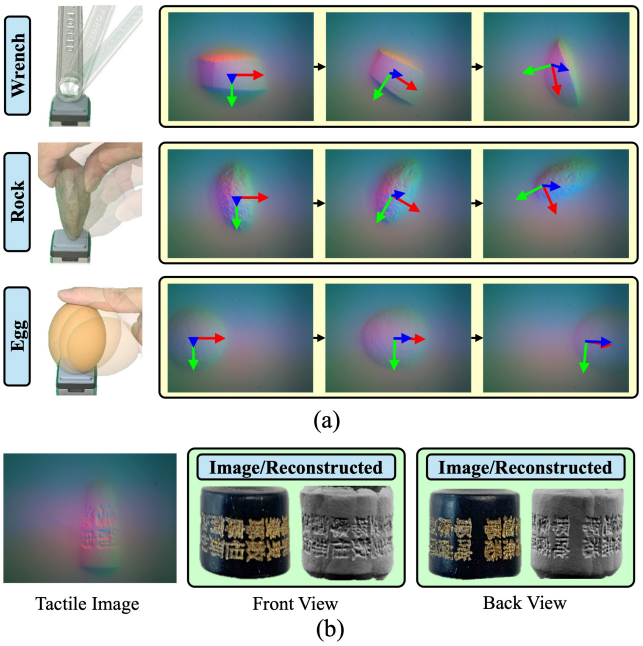

Fig. 1: NormalFlow performs fast, accurate, and robust 6DoF object tracking based on only touch sensing. (a) Accurate tracking of a wide variety of objects, including a wrench, a rock, and even low-texture object like an egg. (b) Applying NormalFlow to tactile-based 3D reconstruction of a 12mm wide bead highlights NormalFlow's high accuracy.

[1]Hung-Jui Huang and Michael Kaess are with Carnegie Mellon University, Pittsburgh, PA, USA {hungjuih, kaess}@andrew.cmu.edu

[2]Wenzhen Yuan is with University of Illinois Urbana-Champaign, Champaign, IL, USA yuanwz@illinois.edu

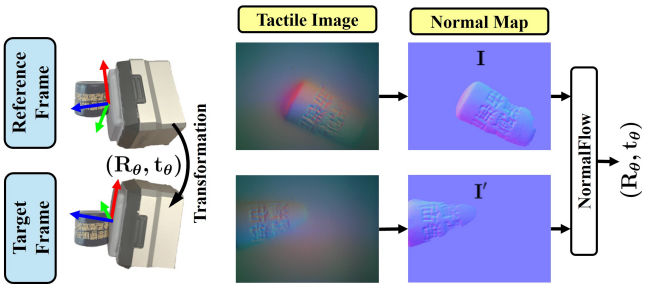

Fig. 2: Given two tactile images before and after object movement, we derive the surface normal maps. NormalFlow determines the object transformations by minimizing discrepancies between the surface normal maps.

Fig. 3: Objects in the tracking experiment.

minimize Eq. (1) iteratively. Linearizing Eq. (1) at the current estimate of $\boldsymbol{\theta}$ results in:

$$\sum_{(u,v)\in\overline{C}} \left[\left(\mathbf{I}'(\mathbf{W}) - \mathbf{R}_{\boldsymbol{\theta}}\mathbf{I}\right) + \left(\nabla\mathbf{I}' \frac{\partial \mathbf{W}}{\partial \boldsymbol{\theta}} - \frac{\partial (\mathbf{R}_{\boldsymbol{\theta}}\mathbf{I})}{\partial \boldsymbol{\theta}}\right) \boldsymbol{\Delta\theta} \right]^2 \quad (2)$$

This linear least squares problem in $\boldsymbol{\Delta\theta}$ is solved in closed form. To improve efficiency, we also adopt the inverse compositional method [9].

NormalFlow offers advantages over ICP by leveraging surface normals for pose estimation. For example, surface tilt can be directly inferred from normal rotations. For a textured ball, ICP aligns global shape, ignoring textures, whereas NormalFlow uses the variation in normal directions to estimate pose from texture.

III. Experiments and Results

We evaluate the tracking accuracy and runtime of NormalFlow on 10 objects (Fig. 3). We collect seven tracking trials for each object, with contact initiated at different poses. Trials average 10.2 seconds in duration.

NormalFlow is compared against three point cloud registration baselines: Point-to-Plane ICP [6], FilterReg [10], and FPFH + RANSAC + ICP [5] (FPFH+RI). Average tracking errors are reported in Table I, and two example trials using NormalFlow are shown in Fig. 4. NormalFlow outperforms all baselines, especially on low-texture objects. FPFH+RI frequently falls into local minima, highlighting the difficulty of extracting reliable features from tactile point clouds. ICP performs consistently worse than both NormalFlow and FilterReg. While NormalFlow only marginally outperforms FilterReg on highly textured objects (e.g., avocado), it shows a clear advantage on less textured ones (e.g., wrench, gammer), maintaining robust tracking where FilterReg often fails. To the extreme, NormalFlow can robustly track objects like a flat table, which is considered textureless by human standards.

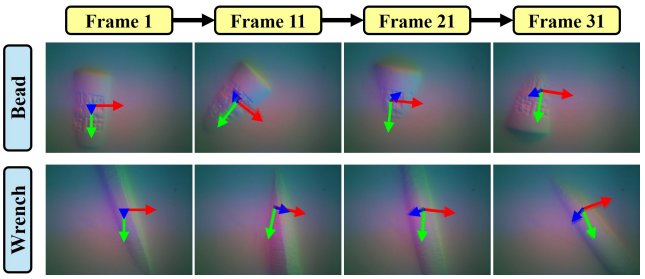

Fig. 4: Example trials on two objects. RGB axes show NormalFlow estimated poses. Transparent RGB axes show true poses, nearly overlapping with NormalFlow poses.

Method	x(mm)	y(mm)	z(mm)	$\theta_x(^\circ)$	$\theta_y(^\circ)$	$\theta_z(^\circ)$
NormalFlow	**0.17**	**0.18**	**0.15**	**1.13**	**1.42**	**0.64**
FilterReg	0.85	1.05	0.20	1.96	2.59	15.4
ICP	1.22	3.44	0.85	2.27	3.30	15.9
FPFH+RI	2.38	1.69	1.26	2.93	36.8	27.8

TABLE I: 6DoF tracking MAE

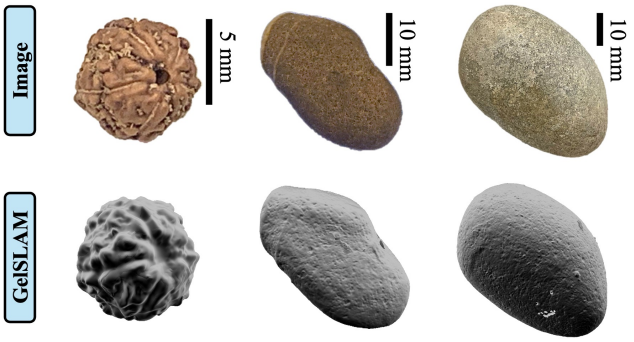

Fig. 5: Reconstruction results from GelSLAM.

On a laptop equipped with an AMD Ryzen 7 PRO 7840U CPU without GPU acceleration, the average runtime of NormalFlow is 13.9 ms, closely comparable to ICP at 13.6 ms, and significantly faster than FilterReg at 145 ms and FPFH+RI at 127 ms.

IV. GelSLAM: Tactile-based 3D Reconstruction

We propose GelSLAM, a real-time, tactile-only 3D reconstruction algorithm built on NormalFlow. In our experiment, the target object is manually rolled across the GelSight Mini, revealing small surface patches in each tactile frame. NormalFlow tracks the 6DoF pose over time, and when a loop closure is detected, it estimates the relative pose between the two endpoints. These poses are optimized in real time using pose graph optimization. As shown in Fig. 5, GelSLAM produces highly detailed reconstructions of the object surface—details that are often difficult to capture with visual methods. Compared to prior approaches such as [2], GelSLAM delivers significantly improved 3D reconstruction quality. Its success highlights the precision of NormalFlow, where even small errors can cause severe artifacts in the final mesh.

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, 2017.

[2] J. Zhao, M. Bauza, and E. H. Adelson, "Fingerslam: Closed-loop unknown object localization and reconstruction from visuo-tactile feedback," 2023.

[3] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess, J. Ortiz, and M. Mukadam, "Neural feels with neural fields: Visuo-tactile perception for in-hand manipulation," 2023.

[4] P. Sodhi, M. Kaess, M. Mukadanr, and S. Anderson, "Patchgraph: In-hand tactile tracking with learned surface normals," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE Press, 2022, p. 2164–2170.

[5] J. Lu, Z. Wan, and Y. Zhang, "Tac2structure: Object surface reconstruction only through multi times touch," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1391–1398, 2023.

[6] Y. Chen and G. G. Medioni, "Object modeling by registration of multiple range images," *Proceedings. 1991 IEEE International Conference on Robotics and Automation*, pp. 2724–2729 vol.3, 1991.

[7] S. Wang, Y. She, B. Romero, and E. H. Adelson, "Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[8] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'81. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981, p. 674–679.

[9] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221 – 255, March 2004.

[10] W. Gao and R. Tedrake, "Filterreg: Robust and efficient probabilistic point-set registration using gaussian filter and twist parameterization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 087–11 096.