

FingerSLAM: Closed-loop Unknown Object Localization and Reconstruction from Visuo-tactile Feedback

Jialiang Zhao^{1,2}, Maria Bauza², and Edward H. Adelson¹

Abstract—In this paper, we address the problem of using visuo-tactile feedback for 6-DoF localization and 3D reconstruction of unknown in-hand objects. We propose FingerSLAM, a closed-loop factor graph-based pose estimator that combines local tactile sensing at finger-tip and global vision sensing from a wrist-mount camera. FingerSLAM is constructed with two constituent pose estimators: a multi-pass refined tactile-based pose estimator that captures movements from detailed local textures, and a single-pass vision-based pose estimator that predicts from a global view of the object. We also design a loop closure mechanism that actively matches current vision and tactile images to previously stored key-frames to reduce accumulated error. FingerSLAM incorporates the two sensing modalities of tactile and vision, as well as the loop closure mechanism with a factor graph-based optimization framework. Such a framework produces an optimized pose estimation solution that is more accurate than the standalone estimators. The estimated poses are then used to reconstruct the shape of the unknown object incrementally by stitching the local point clouds recovered from tactile images. We train our system on real-world data collected with 20 objects. We demonstrate reliable visuo-tactile pose estimation and shape reconstruction through quantitative and qualitative real-world evaluations on 6 objects that are unseen during training.

I. INTRODUCTION

Understanding an in-hand object’s pose and shape is a fundamental component to many robotic manipulation tasks, such as grasping [1], assembly [2], and tool using [3]. While pose estimation and shape reconstruction through vision have been active areas of research in the computer vision domain [4–6], physical interactions through touch provide humans with an effective yet often underestimated way to understand the scene [7]. With the recent advance in high-resolution tactile sensing [8,9], researchers finally have the means to obtain detailed contact information such as the shape and force in a real-time manner. Enabled by these newly developed sensors, several works focus on scene understanding relying solely on high definition tactile sensing [10–16].

However, humans use both vision and tactile during shape inference and localization [17]. Vision provides a global understanding of the scene, while tactile provides a detailed view for the contact geometry. Both modalities are needed for precise localization and reconstruction. In this work, we tackle the problem of localizing and reconstructing an unknown in-hand object by combining the two sensing modalities.

¹Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology {alanzhao, adelson}@csail.mit.edu

²Mechanical Engineering, Massachusetts Institute of Technology bauza@mit.edu

To solve this problem, we need to answer the following question: How can we calculate an optimal pose given multiple sensory estimates that do not always agree with each other? Simultaneous Localization And Mapping (SLAM) is a well-studied problem in the field robotics domain [18]. Central to SLAM is the task of jointly optimizing an objective, often a multi-dimensional pose, by leveraging multiple sensory measurements and producing a more accurate and tractable solution. One popular choice of such optimization frameworks is the factor graph, which models different sensory measurements and pose priors as factors in a bipartite graph. Due to its flexibility and computational efficiency, we chose it as the optimization tool for our visuo-tactile pose estimation and shape reconstruction task.

Our main contributions are:

- A factor graph-optimized 6-DoF pose estimator that combines global vision and local tactile tracking.
- A loop closure matching process that effectively reduces accumulated errors.

We train FingerSLAM with 20 real-world objects, and evaluate it on 6 objects that are unseen during training. FingerSLAM achieves a pose tracking accuracy of $2.8mm/2.4deg$, and produces realistic reconstructed point clouds.

II. RELATED WORKS

Researchers of the robotics community have put forward a wide range of tactile sensing solutions. Sensors working on different sensing principles have been adopted to solve a large set of manipulation tasks. Among different types of tactile sensors, vision-based ones such as GelSight [19] and GelSlim [20] stand out for their rich output, ease to use, and affordability. While we focus on the pose estimation and shape reconstruction task using vision-based tactile sensors, we refer readers to [9] for an in-depth review of different types of tactile sensing and their applications. In this section, we review works on three typical tasks that are most relevant to our solution: slip detection, object property inference, and SLAM.

Slip detection and estimation: Using a similar sensor to ours, Yuan *et al.* compared and analyzed a GelSight tactile sensor’s images collected at different stages of slip in [10] and showed this type of sensors’ capability in detecting micro scale movements. Li *et al.* and Zhang *et al.* trained recurrent neural networks on tactile images to detect slip between multiple time steps in a manipulation sequence [11,21]. Built on their binary slip detection model in [21], Li further added rotational slip direction prediction in [22]. Calandra

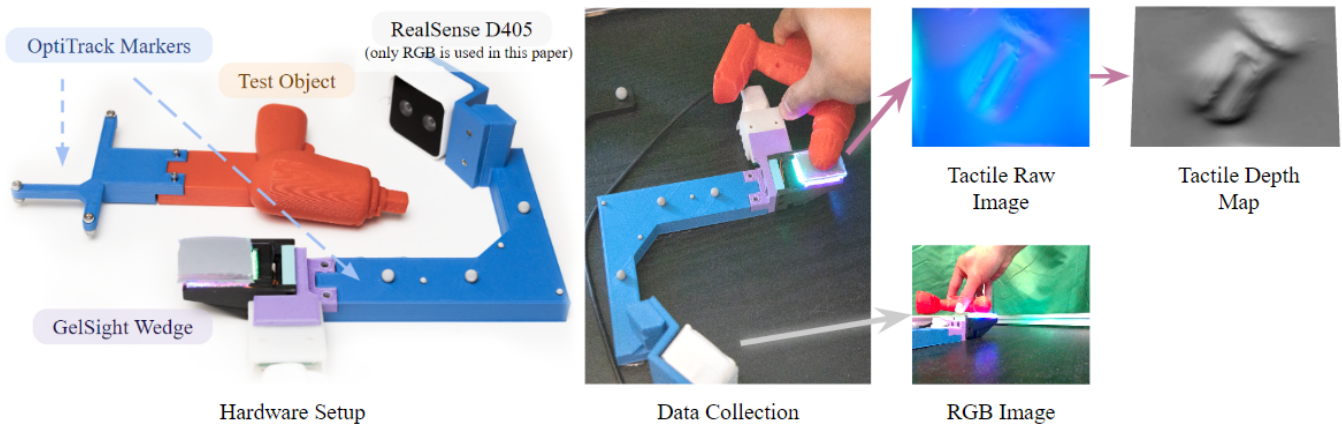


Fig. 1. **Hardware Setup.** We use a GelSight Wedge sensor for tactile sensing, an Intel ReslSense D405 camera mounted on the side for RGB vision sensing, and an OptiTrack setup for motion capture. **Data Collection.** The tactile finger and the camera are fixed to the table at all times. A human operator moves a test object and presses it against the finger. We show sampled tactile and RGB images as well as a reconstructed local tactile depth map on the right.

et al. improved a grasp planner for the classic robot bin-picking problem by incorporating slip detection and achieved a higher grasp success rate [23]. However, those methods only detect slip without localizing the object after the slip. In many precision manipulation tasks we are also interested in the amount of the displacement.

Object property inference and localization: Many works have focused on inferring properties of the in-contact object, such as shape [24–26], texture [27, 28], and material [28–30]. Those learned object properties can be further used for localization. In order to localize current grasps, Bauza *et al.* proposed to match new tactile imprints with previously collected tactile imprints [12], while Luo *et al.* learned to match tactile imprints directly to visual images of the whole object [31]. Assuming known CAD models, Bauza *et al.* proposed to localize by comparing contact masks generated from tactile images with a large bank of random projections of the CAD model [13]. To solve the reverse problem, i.e. what a tactile image looks like given an object and a pose, several tactile simulators have been built to automatically generate tactile images given an object’s CAD model and a finger pose [32, 33]. One major limitation for this category of works is that they all require a known calibrated geometry of the object: a pre-collected tactile map [12], a model of the object [13], or a global image with known geometry [31]. This requirement can be hard to meet in less constraint environments.

Tactile SLAM: Recent studies have shown interests in working with unknown objects by leveraging methods from the SLAM problem. With a focus on 2D shapes, Suresh *et al.* parameterized shapes as Gaussian Process Implicit Surfaces (GPIS), and learned its parameters from tactile signals collected during pushing [14]. Assuming known contact poses, authors of [15] first learned a noisy mapping from known surface geometries to corresponding tactile images, then reconstructed an object by combining many noisy local tactile measurements into an optimized global shape using

factor graph optimization. The closest prior work to ours is [16], where the authors learned to estimate 6D poses and 3D shapes simultaneously for unknown objects. They constructed a pose estimator based on tactile sensing, and a shape reconstruction pipeline that added in new tactile point clouds incrementally on the run. However, this approach heavily relies on the performance of the tactile pose estimator, which lacks a global understanding of the object and can suffer from repeated patterns or smooth surfaces. In contrast, our work combines vision and tactile sensing which provides us with both global and local understandings of the scene without requiring any other domain knowledge. Furthermore, we designed a loop closure mechanism that periodically matches current tactile and vision images to stored key-frames, which significantly reduced accumulated errors. With this, FingerSLAM is able to produce realistic reconstructions even in long sequences.

III. PROBLEM FORMULATION

We consider the problem of estimating an unknown object’s pose while simultaneously reconstructing its shape. Given as inputs a tactile image collected from the finger and a RGB image collected from a wrist-mounted camera, FingerSLAM estimates the object’s 6-DoF pose $P_t \in SE(3)$ and reconstructs its shape as a point cloud S_t incrementally, at each time step t . We show our hardware setup and sample images in Fig. 1.

Assumptions:

- Only one object is present in the same sequence.
- The object is rigid.
- The object is always in contact with the finger, and it is always inside the wrist-mounted camera’s view.

The remainder of this paper is organized with our approach for pose estimation and reconstruction in Sec. IV and the discussion of performance on evaluated objects in Sec. V.

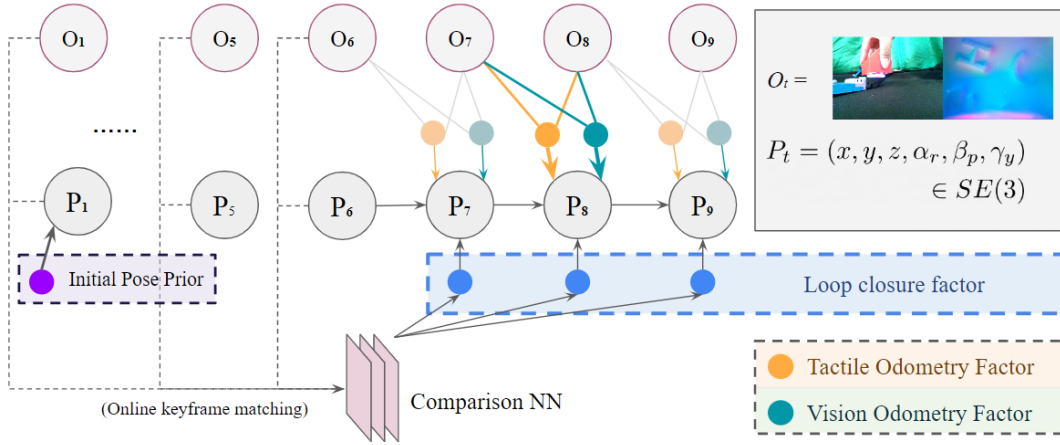


Fig. 2. Factor graph formulation. O_t denotes the observation (tactile and vision images) at each time step, while P_t denotes the estimated $SE(3)$ pose at each time step. Each factor is denoted as color dots.

IV. LOCALIZATION AND RECONSTRUCTION

We begin with pre-processing steps for tactile images, follow with our factor graph formulation of the localization problem, end with our reconstruction approach.

A. Local Surface Reconstruction From Tactile Images

Local surface depth maps are reconstructed with photometric stereo for the GelSight Wedge [8] sensor that we used in this project. On the outside of a GelSight Wedge sensor is a thin layer of elastic silicone gel, which is coated with reflective paint. Once pressed, the gel conforms to the contact surface, and its inner side is illuminated by light sources placed at three sides of the gel with three different colors. A color image is taken, and the color value at each pixel is correlated to the gradients at two directions ∇_x, ∇_y at that location. During post-processing, such images are first unwarped to remove optical distortion. Then the color image is converted to a gradient map using a pre-calibrated look-up table. At each pixel, the gradients ∇_x, ∇_y are recovered. A Poisson solver is used to reconstruct a depth map from this gradient map. An example raw tactile image and its corresponding reconstruction can be found in Fig. 1.

B. Pose Estimation with Factor Graph

Following works in the tactile SLAM domain [14, 16], we choose factor graph to synthesize pose measurements from multiple learned estimators. A factor graph is a bipartite graph, with factors F as constraints and variables P as optimization targets. In our case, P is the object's 6-DoF pose at all time steps, and F consists of a prior F_{pri} and three measurements: tactile odometry F_{tac} , vision odometry F_{vis} , and loop closure F_{lc} . The optimization problem solves for the best $\mathbf{P}^* = P_{t=0:N}^*$ such that its posterior probability is maximized (MAP). Measurement noise for each factor is modeled as a zero-mean Gaussian distribution, with a covariance matrix Σ .

$$\mathbf{P}^* = \arg \max_{\mathbf{P}} \left[\|F_{pri}(P_0)\|_{\Sigma_p}^2 + \sum_{t=1}^N \left(\|F_{tac}(P_{t-1}, P_t)\|_{\Sigma_t}^2 + \|F_{vis}(P_{t-1}, P_t)\|_{\Sigma_v}^2 + \|F_{lc}(P_t)\|_{\Sigma_l}^2 \right) \right]$$

Our factor graph formulation is illustrated in Fig. 2. We used the GTSAM C++ library [34] for factor graph optimization. Next, we discuss details for each factor.

Prior: A unary prior is added to the factor graph only at the initial time step of an episode. This prior anchors the object's estimated pose sequence at a given initial pose. Covariance for the noise model Σ_p is zero.

Tactile odometry with multiple-step refinement: Given tactile depth maps collected at two consecutive time steps that sufficiently overlaps with each other, D_{t-1} and D_t , it is possible to tell the relative movement of the object P_{t-1}^t between the two time steps. We learn the mapping $P_{t-1}^t = f_{tacNN}(D_{t-1}, D_t)$ using convolutional neural networks. The architecture of f_{tacNN} is illustrated on the left of Fig. 3.

Inspired by [4], we propose a multi-pass refinement process which runs f_{tacNN} multiple times until convergence. This process is detailed in Alg. 1. A sequence of resulted intermediate \bar{D}_{t-1}^i as well as D_t are shown in Fig. 4. As i (the number of passes) goes larger, \bar{D}_{t-1}^i appears closer to D_t , which means the estimated pose improves over time.

In experiments we found that the neural network f_{tacNN} learned better when only predicting $SE(2)$ poses, $\{x, y, \gamma_{yaw}\}$, likely due to the fact that the tactile sensor has a flat surface and a relative thin gel, which means the sensor is only sensitive to tangential movements. As such, we only predict $\{x, y, \gamma_{yaw}\}$ with this odometry, and populate $\{z, \alpha_{roll}, \beta_{pitch}\}$ as all zero, and assign large values to the corresponding cells in the covariance matrix Σ_t for the noise model. Values in other cells of Σ_t are chosen empirically.

Vision odometry: Taken input of RGB images captured from a wrist-mounted camera at two consecutive time steps, I_{t-1} and I_t , another CNN f_{visNN} is trained to learn the

Algorithm 1: Tactile pose estimator with multi-step refinement

Data: D_{t-1}, D_t
Result: P_{t-1}^t

$$P_{t-1}^t \leftarrow \begin{bmatrix} I_{3 \times 3} & [0]_{3 \times 1} \\ 0 & 1 \end{bmatrix};$$

$$\bar{D}_{t-1}^0 \leftarrow D_{t-1};$$

 $i \leftarrow 1;$
 $\sigma \leftarrow$ a convergence threshold chosen empirically;

while $i \leq 10$ **do**
 $\Delta P \leftarrow f_{tacNN}(\bar{D}_{t-1}^{i-1}, D_t);$

▷ Run tacNN to obtain a delta estimate

 $\bar{D}_{t-1}^i \leftarrow \text{warp}(\bar{D}_{t-1}^{i-1}, \Delta P);$

▷ Warp \bar{D}_{t-1}^{i-1} with the delta estimate

 $P_{t-1}^t \leftarrow \Delta P \cdot P_{t-1}^t;$

▷ Apply the delta estimate to the result

if $\|\Delta P\| < \sigma$ **then**

▷ Break early if converged

return $P_{t-1}^t;$

end
 $i \leftarrow i + 1;$
end
return $P_{t-1}^t;$

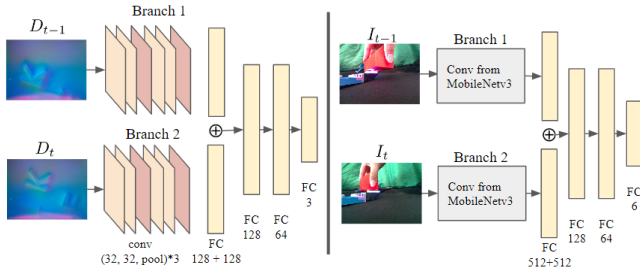


Fig. 3. Neural network architecture for tactile odometry and vision odometry. (left) The tactile odometry network is composed of two convolutional branch models followed by three fully connected (FC) layers. (right) The vision odometry is constructed similarly to the tactile odometry, with the branch model being replaced with convolutional layers extracted from a pre-trained network, MobileNetV3. In both networks, the two branch models share weights.

mapping $P_{t-1}^t = f_{visNN}(I_{t-1}, I_t)$. The architecture is illustrated on the right of Fig. 3. Unlike f_{tacNN} , f_{visNN} predicts the full $SE(3)$ pose because visual images capture global information. We run it for only one pass at each time step. Following examples of many works on visual perception, we bootstrap f_{visNN} with feature layers of models that are pre-trained on a larger image dataset. MobileNetV3 [35] is chosen for this task, due to its compact size. We extract the convolutional layers from the pre-trained model and follow them with four fully connected layers. The pre-trained layers were frozen at the beginning of training, then they were un-frozen to allow for fine-tuning. Noise covariance Σ_v is chosen empirically.

Loop closure: Starting from the 3^{rd} time step, loop closure is checked between each new time step and all previous time steps that are at least 3 steps earlier, i.e. between the

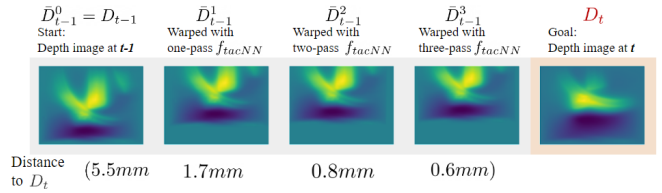


Fig. 4. A sample sequence of intermediate and goal tactile images from multi-step refined pose estimation. D_{t-1} and D_t are the raw tactile images collected at time step $t-1$ and t . Intermediate \bar{D}_{t-1}^i 's are warped versions of D_{t-1} according to the estimated pose at each step. After the last step $i = 3$ in the refinement process, the error was reduced to $0.6mm$, compared to the original movement of $5.5mm$. Visually, \bar{D}_{t-1}^i gets closer to D_t as i increases.

two time steps $t_{i,i>2}$ and $t_{j,j=\forall[1,i-2]}$ we check $LC(i, j) = \{True, False\}$. Two neural networks are trained for this process, $G_{tac}(D_i, D_j)$ and $G_{vis}(I_i, I_j)$. Both output a scalar value between 0 and 1 which represents the likelihood that there is a loop closure between t_i and t_j . $G_{tac}(D_i, D_j)$ takes tactile depth maps as inputs and $G_{vis}(I_i, I_j)$ takes RGB images from the wrist-mounted camera as inputs. Both neural networks share the same architecture as the odometry networks but with a different output layer, whose size is equal to one. Then, we let

$$LC(i, j) = a \times G_{tac}(D_i, D_j) + b \times G_{vis}(I_i, I_j) \stackrel{?}{>} c \quad (1)$$

with a, b as weights and c as a threshold. All three values are chosen empirically.

If any pair of two time steps is considered as a match, we run both the vision odometry and the multi-pass tactile odometry between this pair, and add the measurement to the factor graph.

C. Shape Reconstruction

With tactile depth maps D_t and the optimized object's pose P_t at each time step, we reconstruct the object's shape by stitching all D_i 's together. Since the reconstructed shape is not used for any further prediction, the reconstruction is done trivially by scaling D_t to its true size, converting it to a point cloud, anchoring the point cloud to pose P_t , and finally reducing the whole point cloud with voxel down-sampling. Voxel size is chosen empirically.

V. EXPERIMENTS AND EVALUATION

A. Experiment Setup

We collected real world data (D_t, I_t, P_t) 's for both training and evaluation. In this section we discuss our data collection setup, with the procedure shown in Fig. 1.

Sensors: We used a GelSight Wedge finger for tactile sensing and an Intel RealSense™ D405 camera mounted on the side of the GelSight sensor to mimic the view of a wrist-mounted camera. Note that even though the D405 camera is a near-range RGB-D camera, we found its depth sensing to be rather noisy in our settings, thus only RGB information was used.

Motion Capture: We used 4 OptiTrack Flex-3™ cameras to capture the ground truth movement of the object. We

attached a number of reflective markers to the base of the finger and the base of the object. During experiments, the finger and the RGB camera were always fixed on the desk. A human operator held the object and pressed the object against the finger tip at different locations. Images from both sensors (D_t, I_t) as well as the ground truth pose of the object in the finger’s coordinate frame P_t were recorded at each time step.

Objects: We used 20 3D-printed objects for training, as well as 4 3D-printed and 2 real-world objects for evaluation. The training objects are shown in Fig. 5. Each individual evaluated object will be discussed in Sec. V-B. All objects were selected from the YCB object dataset [36] or the ABC object dataset [37]. The two real-world objects are identical glass jars, except one was painted red and the other one was clear. The two jars were used to test the generalizability of our pose estimator. We collected about 350 data points on each object, which gave us roughly 7,000 raw training data and 2,500 raw evaluation data. Because of the difficulty in collecting a larger amount of data in the real world, we augmented the training data in two ways:

- For the tactile pose estimator, we artificially translate and rotate the depth images randomly. We first generated multiple random $SE(2)$ poses, \hat{p} ’s, for each $(P_{t-1}^t, D_{t-1}, D_t)$ pair, and we applied \hat{p} ’s to P_{t-1}^t and D_{t-1} . The generated $(\hat{P}_{t-1}^t, \hat{D}_{t-1}, D_t)$ ’s were used as augmented training data.
- For both tactile and vision, instead of only using data points collected at consecutive time steps, we also used data points collected at further away time steps as long as the relative displacement between them was smaller than $12mm$.

The final size of our training dataset was 40,000 for tactile and 25,000 for vision. No augmentation was done to the evaluation data.



Fig. 5. Objects used during training. We selected 20 objects from the YCB and the ABC dataset.

B. Pose Tracking Performance

In this section two evaluation methods and their results are discussed: (1) an ablation study of the pose estimator’s per-component performance, and (2) a comparison between the proposed pose estimator’s performance and the size of the raw movement.

Ablation study: Detailed numerical results of the proposed system’s per-component performance on each object

are shown in Tab. I. The *Raw* column shows the median of the raw movement between two time steps. This was calculated from the poses collected from the motion capture system. The rest of the four columns shows the median error obtained from each corresponding pose estimator.







Tactile v. Vision: Comparing the top three objects against the bottom three, we see a clear trend that the tactile estimator performed better when there were unique and fine textures on the object. We found the tactile estimator to perform the worst on the *Mini Bleach* object, likely due to its mostly smooth surface. The vision estimator struggled on the two jars, which we believe was attributed to their significantly larger sizes than all the training objects. Comparing the two jars, tactile performed almost equally, while vision performed worse on the transparent one. This shows that estimating with tactile has the advantage of generalizability across apparent properties.

Factor graph optimization: The *Combined* column shows prediction errors of the optimized estimator which combined tactile and vision, without loop closure detection. For the bottom 4 objects, the *Combined* method achieved a performance in between the performances of tactile standalone and vision standalone. For the top 2 objects, *Combined* outperformed both standalone methods. In those cases, the factor graph was able to retrieve good individual channels and compose a 6-DoF prediction that had a higher accuracy. Moreover, the combined method produced very close accuracy on the two jars, despite the vision estimator’s much poorer performance on the transparent jar. This further confirms that the combined method was able to pick more valuable channels between the estimators.

Closed-loop factor graph: The last column shows the performance of FingerSLAM, the factor graph-optimized estimator with loop closure checking. We did parameter tuning on the loop closure checking threshold. The best result was obtained with $a = 0.3, b = 0.7, c = 0.9$ in Eq. 1. Vision was assigned a higher weight for this loop closure process, because tactile suffered more from repeated patterns. Looking at the *w/ Loop Closure* column, FingerSLAM consistently outperformed the *Combined* method for all objects. In examinations of per-step errors within an episode, we found that loop closure significantly helped cases when a newly visited spot was close to a previously visited spot.

Performance v. step size: A pose estimator’s error hugely depends on the size of the original movement. We grouped all our evaluation data points in three groups according to the size of their original movement: (a) data pairs that has a movement of less than $5mm$, (b) data pairs of movement between $5mm$ and $10mm$, and (c) data pairs with more than $10mm$ movement. Detailed per-component performance for each movement group is shown in Tab. II. When step sizes are small (the $< 5mm$ group), tactile sensing provides a more accurate estimation compared to vision, which we believe attributed to the fact that tactile has a higher resolution for local movements. The situation reversed when the step size was large (the $> 10mm$ group), likely due to the increased difficulty in finding correlations from two tactile

TABLE I
P.E. ERROR PER COMPONENT PER OBJECT

Name	Picture	Raw mm : deg	Tactile only mm : deg	Vision only mm : deg	Combined mm : deg	w/ Loop Closure mm : deg
Letter Cube		6.18 : 6.34	3.10 : 2.79	2.99 : 2.61	2.87 : 2.29	2.27 : 2.18
Graphic		5.93 : 5.59	2.45 : 2.61	2.86 : 2.11	2.28 : 2.13	2.13 : 2.07
Rubik's Cube		5.88 : 5.66	2.33 : 2.80	2.97 : 2.37	2.50 : 2.43	2.30 : 2.41
Mini Bleach		5.95 : 5.74	4.52 : 3.77	3.25 : 2.54	4.06 : 2.91	3.46 : 2.50
Painted Glass Jar		6.40 : 6.21	3.50 : 2.70	4.12 : 2.94	3.63 : 2.63	3.19 : 2.82
Glass Jar		6.46 : 6.02	3.58 : 2.47	4.79 : 4.21	3.79 : 2.27	3.38 : 2.27

images when the overlapping area is small. Averaging all our evaluation data (the *all* row), the tactile estimator performed better than the vision estimator, and our proposed closed-loop combined method achieved the best overall pose estimation result.

TABLE II
P.E. ERROR AT EACH RAW MOVEMENT SIZE GROUP

Raw Group	Raw Median mm:deg	Tactile mm:deg	Vision mm:deg	Combined mm:deg	w/ LC mm:deg
<5mm	3.1 : 3.8	1.0 : 2.3	2.6 : 2.7	1.8 : 2.3	1.4 : 2.3
5-10mm	7.1 : 5.7	3.3 : 2.9	3.9 : 3.0	3.8 : 2.8	2.8 : 2.8
>10mm	11.7 : 5.9	6.8 : 3.1	6.1 : 3.2	6.3 : 3.1	6.0 : 3.1
all	6.7 : 5.6	3.0 : 2.7	3.5 : 2.9	3.2 : 2.4	2.8 : 2.4

C. Reconstruction Performance

Qualitative reconstruction results of two objects are shown in Fig. 6. Each sequence used for reconstruction consists of 21 time steps, with every 7 time steps as an episode. After each episode, the factor graph was reset and a new pose from the motion capture system was used as the initial pose prior for the next episode. Comparing shapes reconstructed from the tactile pose estimator only (middle column) and the proposed FingerSLAM pose estimator (right column), FingerSLAM wins clearly with much less mis-alignments. Comparing shapes reconstructed from OptiTrack poses (left column) and FingerSLAM (right column), we see that the two columns share similar overall shape. The right column shapes are easily recognizable, and appear consistent with the left column despite small mis-alignments in certain areas.

VI. CONCLUSIONS

In this paper, we present an approach to localize and reconstruct an unknown shape through vision and tactile feedback. A factor graph-based framework is proposed to jointly optimize all measurement units, and we close the

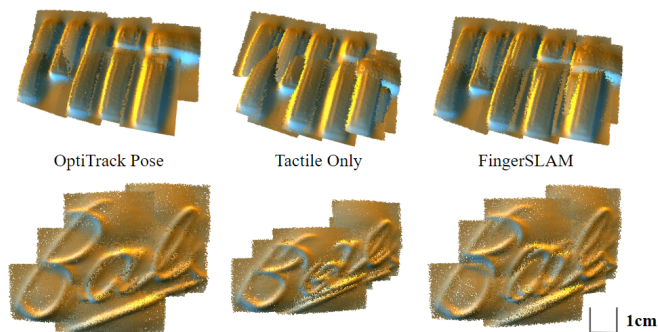


Fig. 6. Shaped reconstructed from 21 local tactile patches, with poses produced by the motion capture system (left), the tactile pose estimator only (middle), and the proposed estimator (right).

loop by actively checking for loop closures. Such a system has the advantage of utilizing both global features and local textures of an object for pose tracking and reconstruction. Quantitative and qualitative evaluations show that our proposed method is able to achieve high tracking accuracy and produce high-fidelity reconstructed shapes on novel objects.

In the future, this approach can be potentially improved by utilizing force vectors at the contact surface, which encapsulate unique higher order information of the movement. Furthermore, following works on neural probabilistic inference [38], instead of using fixed covariance matrices for each factor, having the networks predict both the most likely pose and its variance can potentially help the factor graph make better decisions.

ACKNOWLEDGMENT

Toyota Research Institute provided funds to support this work. The authors thank Dr. Shaoxiong Wang for providing support for the GelSight sensor, Sandra Q. Liu for helping the authors to make suitable gels, and Branden Romero for the support in setting up the motion capture system.

REFERENCES

- [1] C. Ferrari and J. F. Canny, "Planning optimal grasps." in *ICRA*, vol. 3, no. 4, 1992, p. 6.
- [2] J. Zhao, D. Troniak, and O. Kroemer, "Towards robotic assembly by predicting robust, precise and task-oriented grasps," *arXiv preprint arXiv:2011.02462*, 2020.
- [3] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020.
- [4] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3343–3352.
- [5] 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.
- [6] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, et al., "Bop: Benchmark for 6d object pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 19–34.
- [7] R. L. Klatzky, S. J. Lederman, and V. A. Metzger, "Identifying objects by touch: An "expert system"," *Perception & psychophysics*, vol. 37, no. 4, pp. 299–302, 1985.
- [8] S. Wang, Y. She, B. Romero, and E. 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, pp. 6468–6475.
- [9] Z. Kappassov, J.-A. Corrales, and V. Perdereau, "Tactile sensing in dexterous robot hands," *Robotics and Autonomous Systems*, vol. 74, pp. 195–220, 2015.
- [10] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson, "Measurement of shear and slip with a gelsight tactile sensor," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 304–311.
- [11] Y. Zhang, Z. Kan, Y. A. Tse, Y. Yang, and M. Y. Wang, "Fingervision tactile sensor design and slip detection using convolutional lstm network," *arXiv preprint arXiv:1810.02653*, 2018.
- [12] M. Bauza, O. Canal, and A. Rodriguez, "Tactile mapping and localization from high-resolution tactile imprints," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3811–3817.
- [13] M. Bauza, A. Bronars, and A. Rodriguez, "Tac2pose: Tactile object pose estimation from the first touch," *arXiv preprint arXiv:2204.11701*, 2022.
- [14] S. Suresh, M. Bauza, K.-T. Yu, J. G. Mangelson, A. Rodriguez, and M. Kaess, "Tactile slam: Real-time inference of shape and pose from planar pushing," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 322–11 328.
- [15] S. Suresh, Z. Si, J. G. Mangelson, W. Yuan, and M. Kaess, "Shapemap 3-d: Efficient shape mapping through dense touch and vision," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7073–7080.
- [16] P. Sodhi, M. Kaess, M. Mukadam, and S. Anderson, "Patchgraph: In-hand tactile tracking with learned surface normals," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2164–2170.
- [17] R. L. Klatzky, S. J. Lederman, and C. Reed, "There's more to touch than meets the eye: The salience of object attributes for haptics with and without vision." *Journal of experimental psychology: general*, vol. 116, no. 4, p. 356, 1987.
- [18] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.
- [19] 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.
- [20] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, "Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1927–1934.
- [21] J. Li, S. Dong, and E. Adelson, "Slip detection with combined tactile and visual information," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7772–7777.
- [22] J. Li, "Rotational direction detection using tactile sensor and external camera," in *2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2019, pp. 1456–1461.
- [23] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" *arXiv preprint arXiv:1710.05512*, 2017.
- [24] C. Strub, F. Wörgötter, H. Ritter, and Y. Sandamirskaya, "Using haptics to extract object shape from rotational manipulations," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 2179–2186.
- [25] S. Luo, X. Liu, K. Althoefer, and H. Liu, "Tactile object recognition with semi-supervised learning," in *International Conference on Intelligent Robotics and Applications*. Springer, 2015, pp. 15–26.
- [26] S. Luo, W. Mou, K. Althoefer, and H. Liu, "iclap: Shape recognition by combining proprioception and touch sensing," *Autonomous Robots*, vol. 43, no. 4, pp. 993–1004, 2019.
- [27] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2722–2727.
- [28] W. Yuan, S. Wang, S. Dong, and E. Adelson, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5580–5588.
- [29] O. Kroemer, C. H. Lampert, and J. Peters, "Learning dynamic tactile sensing with robust vision-based training," *IEEE transactions on robotics*, vol. 27, no. 3, pp. 545–557, 2011.
- [30] E. Kerr, T. M. McGinnity, and S. Coleman, "Material recognition using tactile sensing," *Expert Systems with Applications*, vol. 94, pp. 94–111, 2018.
- [31] S. Luo, W. Mou, K. Althoefer, and H. Liu, "Localizing the object contact through matching tactile features with visual map," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 3903–3908.
- [32] Z. Si and W. Yuan, "Taxim: An example-based simulation model for gelsight tactile sensors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2361–2368, 2022.
- [33] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra, "Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3930–3937, 2022.
- [34] F. Dellaert, "Factor graphs and gtsam: A hands-on introduction," Georgia Institute of Technology, Tech. Rep., 2012.
- [35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
- [36] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*. IEEE, 2015, pp. 510–517.
- [37] S. Koch, A. Matveev, Z. Jiang, F. Williams, A. Artemov, E. Burnaev, M. Alexa, D. Zorin, and D. Panozzo, "Abc: A big cad model dataset for geometric deep learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [38] J. Zhao, J. Liang, and O. Kroemer, "Towards precise robotic grasping by probabilistic post-grasp displacement estimation," in *Proc. Field Service Robot*. Springer, 2019.