State Estimation - Literature Review

Haniyeh Altafi
haniyeh.altafi@torontomu.ca

***Abstract*—**

I. Introduction

Haptic SLAM and Neural Radiance Fields (NeRF) are two cutting-edge technologies that are transforming the field of robotics and computer vision. Haptic SLAM refers to the simultaneous localization and mapping of objects and environments using touch, while NeRF is a deep learning model that enables the rendering of 3D scenes from 2D images. The combination of these technologies opens up new opportunities for robots to interact with their surroundings in a more natural and human-like manner, overcoming the limitations of individual modalities. For example, robots can use haptic SLAM to generate a map of an environment and use NeRF to render virtual views of objects and environments. This enables robots to perceive and interact with objects and environments in a more comprehensive manner, overcoming the limitations of individual modalities.

Haptic SLAM is crucial for robots to interact with objects and environments in a safe and controlled manner. Unlike visual SLAM, which relies on cameras to perceive the environment, haptic SLAM uses touch sensors to sense the physical properties of objects and the environment. This is particularly important in scenarios where vision is occluded or unreliable, such as in cluttered or poorly lit environments. By combining sensory information from touch, robots can estimate the position and orientation of objects and build up a map of the environment. This information can then be used to guide robot manipulation tasks, such as grasping and manipulating objects [1].

NeRF, on the other hand, is a deep learning model that enables robots to understand the 3D structure of objects and environments from 2D images. Given a set of RGB-D images of a scene, NeRF can generate high-quality RGB-D images for novel view orientations of the scene. This opens up new possibilities for robots to interact with objects and environments in ways that were previously not possible. For example, robots can use NeRF to generate virtual views of objects, enabling them to perceive and interact with objects from any angle, even if they are not directly visible [2].

II. Problem Statement

The aim of this project is to determine the location and orientation of an object through tactile perception. By physically touching the object, we can obtain detailed information about its properties, especially when the robot's view is obstructed. Recent advancements in touch sensors enable precise measurements at the point of contact. However, a single measurement alone cannot provide a complete understanding of the object's position; multiple measurements need to be analyzed together to address this inference problem. By utilizing all the measurements, we can derive a sequence of potential object positions.

Tactile perception plays a crucial role in robotics applications, particularly in manipulation tasks. However, collecting tactile data is more time-consuming compared to visual data, which limits its integration into machine learning solutions for robotics.

The novelty of my research lies in using tactile images generated by the Digit sensor to build a deep learning model with NeRF. This approach is distinct from previous studies that solely relied on visual images. By incorporating touch sensors, we can expand the capabilities of NeRF and enable diverse applications in fields such as medical imaging, biomedical applications, and remote surgery.

The integration of tactile sensory data into the NeRF framework enhances its capabilities and unlocks new possibilities. This pioneering effort in employing touch sensors in NeRF models has the potential to revolutionize fields like medical imaging, where it can provide a comprehensive understanding of anatomical structures and facilitate accurate diagnostics. Additionally, it can benefit biomedical applications by enabling more precise studies of object-human touch interactions. In remote surgery, tactile sensing integrated into NeRF models can enhance surgeons' perception by providing real-time haptic feedback.

This research aims to demonstrate the feasibility and effectiveness of utilizing tactile images in NeRF models and showcase their potential impact on medical imaging, biomedical applications, and remote surgery. By expanding the range of sensory inputs and incorporating touch, we can empower researchers and practitioners with enhanced tools for analysis, diagnosis, and surgical procedures.

Overall, this research bridges the gap between tactile sensing and deep learning, pushing the boundaries of 3D scene reconstruction and object modeling. The use of selective tactile images in NeRF allows for a more efficient process of generating a detailed representation of the object without capturing haptic images for its entire geometry. By leveraging NeRF technology, we can render a complete model using only a subset of tactile images. This innovative methodology streamlines the process and improves efficiency.

III. Literature Review

The literature review section aims to provide a comprehensive and critical analysis of the existing research and scholarly works related to NeRF and Haptic datasets. This review delves into a thorough examination of the literature on NERF, discussing key papers and studies. Each reviewed paper is presented with its title, and main findings, allowing for a comprehensive understanding of the research landscape. Additionally, the methodologies and approaches used in the reviewed papers are analyzed to identify common themes, trends, or challenges in the literature.

By comparing and integrating the findings from NeRF and haptic feedback research, this section sheds light on the potential benefits and limitations of these technologies. The literature review section serves as a foundation for future research, suggesting potential areas of focus and offering insights into the evolving landscape of NeRF and haptics.

A. Utilizing NeRF for Generating Tactile Sensory Data

Zhong et al. [8] present a method for generating realistic tactile sensory data for use in robotics applications. The process starts by using a Neural Radiance Fields (NeRF) model to obtain images of objects of interest from easily obtained camera images. These NeRF-rendered RGB-D images are then used as inputs to a conditional Generative Adversarial Network model (cGAN) to generate tactile images from desired orientations. The generated data is evaluated using metrics such as the Structural Similarity Index and Mean Squared Error, as well as through a tactile classification task in simulation and in the real world. Results indicate that the generated data can improve classification accuracy by around 10% and the model is able to transfer to different tactile sensors with a small fine-tuning dataset.

The article highlights the importance of tactile sensing in human and robotic tasks such as object recognition and manipulation. While tactile data is crucial for these tasks, collecting it is a challenging and time-consuming process. This is because robots need to physically interact with the environment and objects, and the output from different tactile sensors can differ significantly, limiting the validity of the collected data. The lack of standards at the hardware level also contributes to this challenge. The article highlights the need for solutions to overcome these challenges and make it easier to collect and use tactile data for machine learning-based tasks [8].

The article also discusses the challenge of collecting tactile data for use in machine learning tasks and highlights recent approaches that aim to overcome this challenge by generating synthetic tactile sensor responses from easier-to-collect data acquired using different modalities, such as RGB-D images. These vision-based generative approaches have the limitation of requiring visual samples at specific positions and orientations, but recent advances in neural volume-rendering techniques, such as NeRF, have made it possible to synthesize RGB-D images for new view orientations with only sample 2D images and camera poses. This provides additional information on the structure of the scene, which is leveraged to generate tactile images for 3D objects [8].

The paper presents a framework for generating synthetic tactile data from RGB-D images rendered by NeRF models. This is the first method to use a deep generative model conditioned on RGB-D inputs for generating tactile data. The framework eliminates the need for accurate hand-engineered modeling and calibration of the tactile sensor and objects and can handle different 3D object geometries. The use of NeRF models allows for generalization to new object views and eliminates the need for a new input image each time a new object reading is desired. The framework can transfer to new sensors with a small fine-tuning dataset, overcoming the domain gap between different tactile sensors and leveraging different tactile datasets for better performance [8].

In Zhong's paper, a framework for generating synthetic tactile data from RGB-D images is presented. The framework uses NeRF models to obtain images of objects and a cGAN model to generate tactile images from desired orientations. The results show that the generated data can improve tactile classification accuracy and can be transferred to different tactile sensors with a small fine-tuning dataset.

However, the approach has limitations, such as the need for a NeRF model to be trained for each object and the assumption of operating only on rigid objects with a fixed force range and orientation. Generating tactile images for 3D objects and generalizing to different forces and orientations, as well as soft objects, is still an open problem in the tactile-sensing community and is left for future work [8].

B. Learning Tactile Models for Factor Graph-based Estimation

This paper presents a novel approach to estimating object poses from touch during planar pushing using vision-based tactile sensors. The authors propose to directly learn tactile observation models that predict the relative pose of the sensor given a pair of tactile images. This approach allows them to incorporate tactile measurements into a factor graph and solve the inference problem of estimating the latent object state. They demonstrate that their method reliably tracks object poses for over 150 real-world planar pushing trials using tactile measurements alone. [3]

The proposed approach consists of two stages: learning tactile observation models and integrating them into a factor graph. In the first stage, a convolutional neural network (CNN) is trained to predict the relative pose of the sensor given a pair of tactile images. A novel data augmentation technique is proposed to generate more training data. In the second stage, the learned tactile observation models are integrated into a factor graph to solve the inference problem of estimating the latent object state. The authors show that their approach outperforms existing methods that use low-dimensional force measurements or engineered functions to interpret tactile measurements. [3]

The authors propose using a factor graph to incorporate tactile measurements into the estimation of object poses during planar pushing tasks. A factor graph is a graphical model that represents the joint probability distribution over a set of variables and their relationships. In this case, the variables are the latent object poses and the tactile measurements, while the relationships are the physics and geometric constraints that govern the system's behavior. [3]

To incorporate tactile measurements into the factor graph, the authors adopt a two-stage approach. In the first stage, they train local tactile observation models that predict the relative sensor poses given a pair of tactile images. These models are trained using ground truth data and a novel data augmentation technique. In the second stage, the authors integrate these models, along with physics and geometric factors, within a factor graph optimizer. [3]

The factor graph includes three types of factors: tactile factors, physics factors, and geometric factors. Tactile factors predict the relative sensor poses from tactile images, while physics factors model quasi-static pushing, and geometric factors model object and end-effector intersections. The authors also include global pose priors as unary factors on the end-effector and the first object pose. [3]

By incorporating tactile measurements into the factor graph, the authors can reason over a stream of measurements, incorporating structural priors in an efficient, real-time manner. The factor graph enables them to solve the inference problem of estimating the latent object state and track object poses during planar pushing tasks using tactile feedback alone. The authors demonstrate that their approach outperforms existing methods that use low-dimensional force measurements or engineered functions to interpret tactile measurements. [3]

One limitation of this work is its focus on planar pushing tasks. Although reliable object tracking using tactile feedback is demonstrated for planar pushing sequences, it is unclear how well this approach would generalize to other manipulation tasks or objects with different shapes and dynamics. The proposed method may not be directly applicable to scenarios involving more complex manipulation tasks or objects with non-planar surfaces. [3]

Additionally, the paper lacks a comprehensive analysis of the computational efficiency of the proposed approach. While the authors mention the use of a factor graph optimizer, they do not discuss the computational requirements or potential scalability issues when applying this method to larger-scale or real-time applications. Understanding the computational trade-offs and potential limitations of implementing this approach in practical robotic systems would be valuable. [3]

As future work, the authors suggest learning to model a distribution over relative poses instead of only mean-squared error values. They also propose learning richer feature descriptors that describe contact patch geometry in addition to the patch centers. Finally, they suggest incorporating different physics priors to make these tactile factors applicable to more complex manipulation tasks. [3]

C. In-Hand Tactile Tracking with the Learned Surface Normals

Sodhi et al. addressed the problem of tracking 3D object poses during in-hand manipulations using touch. Their focus is on tracking small objects with the use of vision-based tactile sensors that provide high-dimensional tactile image measurements at the point of contact. Unlike previous work, which relied on a priori information about the object being localized, this approach removes that requirement. The key insight is that an object is composed of several local surface patches, each of which is informative enough to allow for reliable object tracking. The geometry of the local patch can also be recovered online by extracting local surface normal information from the tactile image.

A novel two-stage approach is proposed, where the first stage involves learning a mapping from the tactile images to the surface normals using an image translation network, and the second stage involves using the surface normals within a factor graph to reconstruct a local patch map and infer 3D object poses. Reliable object tracking is demonstrated for over 100 contact sequences across unique shapes with four objects in simulation and two objects in the real world [4].

A novel method for tracking the 3D poses of objects from tactile image sequences during in-hand manipulations has been introduced. The method is based on a factor graph and has shown to be reliable in tracking four simulated objects and two real objects without any prior information about the objects. The success of the approach is attributed to its ability to treat a complex object as a combination of local patches that can be mapped and tracked independently, as well as the localization of surface normal information within the tactile image, which is independent of the global object shape [4].

A potential limitation of the approach is the lack of distinguishability in the local patch map, such as flat or featureless patches, or indistinct patch motions in the observed image space, like the rotations of a spherical object. Future work will focus on addressing these challenges by exploring solutions that consider geometric degeneracies and detecting slip and shear to resolve motion degeneracies. The tracker will also be supplemented with a global first pose relocalization that can be applied to objects of various shapes, using visual images to predict the likelihood of contact location [4].

D. A database of textures, utilizing haptic feedback

The vibrations generated while a rigid tool is moved across an object's surface can be used to determine the texture of the surface. These vibrations, which are produced by the interaction between the tool and the surface, can be measured using an accelerometer. However, the temporal and spectral properties of the signals obtained are strongly influenced by factors such as the applied force and the speed at which the tool moves over the surface. Currently, there are no robust features that can be used for texture classification and recognition with varying scan-time parameters.

Strese et al. presents a haptic texture database that allows for a systematic analysis of potential features. The database, which is publicly available, contains recordings of accelerations during controlled scans and human-led, uncontrolled explorations of 43 different textures. To test potential features, we examine six well-established features used in audio and speech recognition, using a Gaussian Mixture Model-based classifier on our recorded free-hand signals. The results show that the best results are obtained using Mel-Frequency Cepstral Coefficients (MFCCs), with a texture recognition accuracy of 80.2% [5].

E. Real-Time Inference of Object Shape and Pose using Planar Pushing and Tactile SLAM

Suresh et al. present a novel method for online shape and pose estimation of a planar object using tactile exploration via contour following. The method is designed to work with a planar pusher-slider system, where a robot manipulator pushes along a planar object while recording a stream of tactile measurements. Their goal was to build a shape contour in real-time as a Gaussian process implicit surface and optimize for pose via geometry and physics-based constraints. [6].

The paper introduces the Gaussian process implicit surface (GPIS) as a nonparametric shape representation that satisfies the requirements of faithfully approximating arbitrary geometries and being amenable to probabilistic updates. The GPIS builds an implicit surface shape representation that is the zero level-set of a GP potential function, and spatial partitioning with local GPs enables efficient regression. The factor graph illustrates the relationship between the variables to be optimized for and the factors that act as constraints. [6].

The system estimates the 2-D shape and object's planar pose in real-time from a stream of tactile measurements. The paper expands the scope of the batch-SLAM method by Yu et al. [7] with a more meaningful shape representation. The method is evaluated in simulated and real experiments, and the results show that it can accurately estimate the shape and pose of different objects in planar pushing tasks. [6].

Overall, the paper presents a significant contribution to the field of tactile exploration and manipulation, as it provides a method for real-time shape and pose estimation of planar objects using tactile sensing. The GPIS representation and factor graph optimization enable efficient and accurate inference, and the method is shown to work well in both simulated and real-world experiments. [6].

IV. Conclusion

References

[1] Feryal MP Behbahani, Guillem Singla-Buxarrais, and A Aldo Faisal. Haptic slam: An ideal observer model for bayesian inference of object shape and hand pose from contact dynamics. In *Haptics: Perception, Devices, Control, and Applications: 10th International Conference, EuroHaptics 2016, London, UK, July 4-7, 2016, Proceedings, Part I 10*, pages 146–157. Springer, 2016.

[2] Yen-Chen Lin, Pete Florence, Andy Zeng, Jonathan T Barron, Yilun Du, Wei-Chiu Ma, Anthony Simeonov, Alberto Rodriguez Garcia, and Phillip Isola. Mira: Mental imagery for robotic affordances. In *6th Annual Conference on Robot Learning*, 2022.

[3] Paloma Sodhi, Michael Kaess, Mustafa Mukadam, and Stuart Anderson. Learning tactile models for factor graph-based estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13686–13692. IEEE, 2021.

[4] Paloma Sodhi, Michael Kaess, Mustafa Mukadanr, and Stuart Anderson. Patchgraph: In-hand tactile tracking with learned surface normals. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2164–2170. IEEE, 2022.

[5] Matti Strese, Jun-Yong Lee, Clemens Schuwerk, Qingfu Han, Hyoung-Gook Kim, and Eckehard Steinbach. A haptic texture database for tool-mediated texture recognition and classification. In *2014 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE) Proceedings*, pages 118–123. IEEE, 2014.

[6] Sudharshan Suresh, Maria Bauza, Kuan-Ting Yu, Joshua G Mangelson, Alberto Rodriguez, and Michael Kaess. Tactile slam: Real-time inference of shape and pose from planar pushing. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11322–11328. IEEE, 2021.

[7] Kuan-Ting Yu, John Leonard, and Alberto Rodriguez. Shape and pose recovery from planar pushing. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1208–1215. IEEE, 2015.

[8] Shaohong Zhong, Alessandro Albini, Oiwi Parker Jones, Perla Maiolino, and Ingmar Posner. Touching a nerf: Leveraging neural radiance fields for tactile sensory data generation. In *6th Annual Conference on Robot Learning*, 2022.