Enhancing Robotic Ultrasound Imaging with ElasticFusion SLAM and Pose Graph Optimization

Shiva Surya Lolla[1]
Department of Robotics Engineering
Worcester Polytechnic Institute
Worcester, Massachusetss 01609
Email: slolla@wpi.edu

***Abstract*—This project advances the localization and 3D reconstruction capabilities in robotic ultrasound (US) imaging by integrating the ElasticFusion SLAM framework with a robotic US system equipped with a stereo RGB-D camera setup. Tailored for RGB-D cameras, ElasticFusion excels in low-texture environments typical of human skin and phantom materials used in US imaging and effectively manages the complex, short loopy trajectories of the probe. To refine pose accuracy, Pose Graph Optimization was applied, integrating outputs from the dual-camera setup. The system's efficacy was validated using the EVO evaluation package against ground truth trajectories from a Franka Emika Panda robot, employing metrics such as Absolute Pose Error (APE) and Relative Pose Error (RPE). Results demonstrated a notable reduction in trajectory errors, with RPE improving from 6mm to as low as 1mm with Sim(3) alignment, and a 50% reduction in ground truth data needs. Additionally, a comparative analysis with ORB SLAM 3, a feature-based SLAM method, showed that ElasticFusion provided superior performance in this application context. However, trajectory accuracy experienced some deterioration when using Pose Graph Optimization. These findings underline the advantages of ElasticFusion in enhancing diagnostic imaging by significantly improving trajectory precision and reducing dependency on extensive ground truth data.**

I. Introduction

Medical imaging plays a pivotal role in allowing physicians to visualize a patient's internal anatomy for diagnosis and ongoing monitoring. Among the various techniques available, ultrasound imaging stands out due to its non-invasive, rapid, safe, and cost-effective nature. Despite these benefits, traditional 2D ultrasound imaging presents several challenges, including the laborious nature of analysis, dependence on operator skill for accurate organ volume measurement, and the risk of erroneous diagnostic conclusions [1]. In response, three-dimensional (3D) ultrasound has emerged as a preferred option. Offering superior spatial resolution and real-time 3D visualization, 3D ultrasound leverages a series of 2D images combined with corresponding probe poses to generate comprehensive 3D representations [1], [2].

The quality of three-dimensional ultrasound (3D US) is critically dependent on the precision of probe pose estimates. To address this, various methods have been developed to achieve highly accurate pose measurements. Traditional techniques often employ electromagnetic (EM) devices and optical trackers to determine the position of ultrasound probes [3], [4]. However, these methods require specialized equipment that not only increases costs but also restricts the probe to the operational field of the devices.

To circumvent these constraints, visual SLAM (vSLAM) strategies have been introduced, utilizing a camera to localize the probe and track its successive poses. Sun et al. [5] developed a vSLAM system for freehand 3D US, which captures skin features in the ultrasound scan area using a monocular camera and calculates optimal probe poses within a Bayesian probabilistic framework. Their study also included 3D US reconstruction using these poses and validated the results through comparisons with actual ultrasound scans. Building on this, Ito et al. [6] enhanced pose accuracy by tracking features on the skin surface and applying camera pose estimation techniques through Structure from Motion (SfM) with a monocular camera. While the aforementioned vSLAM approaches offer promising results, they also present significant limitations: (i) they are implemented in freehand 3D US systems that do not operate in real-time and require post-processing of camera video sequences to determine pose outputs; (ii) the generation of groundtruth data for system validation necessitates additional equipment. Specifically, Sun et al. [5] utilized an optical tracking system to obtain their groundtruth poses, whereas Ito et al. [6] employed a 3D mesh model, measured with a laser scanner, as the groundtruth for assessing localization accuracy through the analysis of reconstructed 3D point clouds.

Monocular and stereo camera SLAM systems often encounter challenges with depth estimation and dense reconstruction in low-texture environments. In response, RGB-D cameras, which provide both color and depth information, have become increasingly favored for overcoming these limitations in SLAM applications, especially for indoor scene reconstruction [7]. A review of the literature within the medical imaging domain reveals a variety of vSLAM systems in use. A notable example includes Qin et al. [8], who utilized ORB-SLAM with a stereo camera for localizing a handheld Optical Coherence Tomography (OCT) probe. However, ORB-SLAM, a feature-based method, faces difficulties in low-texture areas, a challenge also highlighted by Turan et al. [9]. They observed ORB-SLAM's inferior performance compared to their proposed RGB-D SLAM approach, inspired by Whelan et al. [10] and Newcombe et al. [11], in the context of capsule endoscopy. They attributed the better performance of their approach to

its reliance on direct SLAM techniques, which utilize joint photometric-geometric pose estimation. This method proves advantageous in endoscopic imaging, where specularity, noise, and the scarcity of robustly identifiable features in the environment significantly reduce feature matching accuracy across frames.

This project aims to achieve precise localization through RGB-D SLAM, utilizing a stereo RGB-D camera system mounted on a robotic ultrasound probe, and to subsequently perform 3D ultrasound reconstruction from the associated 2D images. Insights from the literature review suggest that the ElasticFusion SLAM system, developed by Whelan et al. [10], is particularly well-suited for robotic ultrasound applications. This suitability is largely due to the low-texture nature of the environments typically encountered in ultrasound imaging, such as human skin and phantom materials, and the complex, short-loop trajectories characteristic of the ultrasound probe's movement. The main contributions of this project include:

1. Development of an end-to-end pipeline for transforming 2D ultrasound images and their corresponding camera poses into a 3D ultrasound reconstruction.
2. Enhanced performance of ElasticFusion with slower camera speeds, with minimal impact from reduced texture.
3. Significant improvement in localization accuracy, where the Mean Relative Pose Error (RPE) ranges from 3.5 to 6 mm using ElasticFusion, which is four times better than that observed with ORB-SLAM3 in 3D ultrasound applications.
4. Post-Sim(3) alignment, a reduced Mean RPE of 1 mm across datasets, indicating consistent rotation, translation, and scale drift between the groundtruth and SLAM output poses.
5. A uniform scaling factor of 0.5 observed across all datasets, enhancing pose accuracy.
6. Reduction of the need for groundtruth data by 50% for Sim(3) alignment while maintaining an Absolute Pose Error (APE) of 3 mm.
7. While Pose Graph Optimization (PGO) increases RPE, future refinements could be made by exploring the underlying causes of performance degradation.
8. Demonstrated the potential of the existing framework in practical applications through successful 3D ultrasound reconstructions using SLAM pose outputs.

II. Methods

ElasticFusion is a real-time RGB-D SLAM system that uses dense frame-to-model camera tracking and windowed surfel-based fusion, along with frequent model refinements through local and global loop closures.

The pipeline from running ElasticFusion to obtaining the resulting 3D ultrasound reconstruction via the estimated trajectories is illustrated in Figure 1.

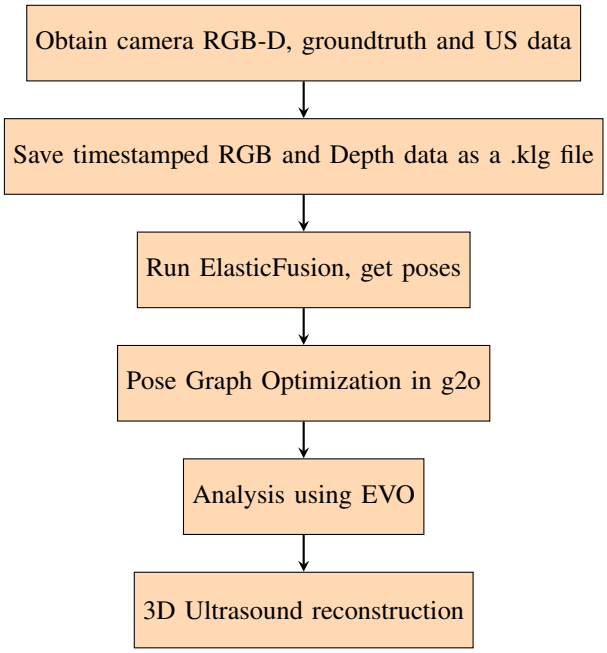


Fig. 1: Pipeline for processing RGB-D and ultrasound data.

Fig. 2: Setup to obtain data for RGB-D SLAM and 3D US reconstruction

A. Obtain camera RGB-D, groundtruth and US data

The RGB data, depth data, groundtruth data and ultrasound images along with their corresponding timestamps are recorded into a rosbag file during the motion of the ultrasound probe attached to the end effector of the robot arm. The setup is shown in Figure 2.

The RGB-D data comprises image sequences from stereo RGB-D cameras mounted on the ultrasound (US) probe. The ground truth data consists of the internal odometry from the Franka Emika Panda robot utilized in this setup. Additionally,

the ultrasound image data includes images captured during the probe's traversal, sourced from the ultrasound machine within the setup.

B. Save timestamped RGB and Depth data as a .klg file

The project utilizes Docker to facilitate the processing of bag files from ROS1, as these are not directly compatible with ROS2. The process is fully automated to ensure efficiency and reproducibility. Here is a simplified overview of the process:

- **Initial Setup:** A Docker container is built from the `ros:noetic-ros-core-focal` image, which is configured to handle ROS-based applications. This container includes essential ROS packages and Python tools necessary for data processing.
- **Data Preparation:** The data is organized into specific directories. These steps are crucial for subsequent processing and analysis.
- **Running the Pipeline:**
  - The Docker container is run with necessary volume bindings to ensure data from host directories is accessible within the container.
  - Multiple ROS commands are executed to playback bag files, extract RGB, depth and US images, and groundtruth poses, and save them with specific formats and resolutions suitable for ElasticFusion.
- **Image and Data Handling:**
  - Saved images are adjusted for ownership and backup.
  - Essential data files such as timestamps and associations are generated to facilitate further processing.
- **Final Steps:** The processed data is converted into a format (.klg) compatible with ElasticFusion.

C. Run ElasticFusion, Get Poses

With the data prepared in the required .klg format, ElasticFusion is executed using the intrinsic parameters for each camera. The trajectory poses for each camera are then saved as text files in the same directory as the .klg file.

D. Pose Graph Optimization in g2o

To refine the pose outputs obtained from running ElasticFusion on a single camera, Pose Graph Optimization (PGO) is employed. The formulation for this optimization is illustrated in Figure 3.

The setup includes two cameras: the main camera, referred to as **Cam 1**, provides the initial pose outputs, and the secondary camera, **Cam 2**, refines these poses within the PGO. Poses from **Cam 1** are denoted by **X**, and those from **Cam 2** by **Z**. An information matrix Ω_i associated with **Cam 2** poses is derived by analyzing their covariance with respect to ground truth data, obtained using the Relative Pose Error (RPE) command from the EVO package. The error between relative poses of **Cam 1** and **Cam 2** is indicated with e_i.

The goal is to find the state x^* which minimizes the error given all measurements:

$$x^* = \arg\min_x \sum_i \mathbf{e}_i^T(x)\Omega_i\mathbf{e}_i(x) \tag{1}$$

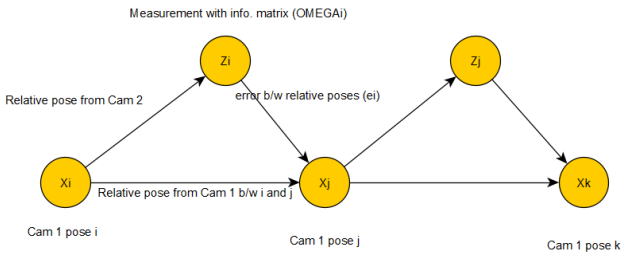


Fig. 3: PGO formulation

The optimization process is outlined in Algorithm 1, wherein input poses undergo continuous refinement, yielding refined poses as the output.

Algorithm 1: PGO optimization algorithm

Result: Converged state x

Function `optimize(`x`)`**:**
  while *not converged* **do**
    (H, b) = buildLinearSystem(x)
    Δx = solveSparse($H\Delta x = -b$)
    $x = x + \Delta x$
  end
  return x

E. Analysis using EVO

In the evaluation of the pose estimation algorithms, two primary metrics are employed using the EVO toolkit: Absolute Pose Error (APE) and Relative Pose Error (RPE).

a) Absolute Pose Error (APE): APE measures the absolute discrepancies between the estimated trajectory and the ground truth trajectory. It is a direct measure of accuracy, evaluating how close the estimated positions at different timestamps are to their true values. The command `evo_ape` is utilized to compute this metric, plot the results, align the trajectories, optionally correct the scale, and save the statistical outcomes in a zip file for further inspection.

b) Relative Pose Error (RPE): RPE assesses the local accuracy of the trajectory over a specified segment length or time interval, focusing on the translational part (`trans_part`) of the motion. This metric provides insights into translational drift every 5mm, highlighting how positional differences evolve over time and indicating the local tracking performance. Commands for RPE analysis include options for trajectory alignment, scale correction, and result visualization. **This project aims to obtain an RPE of 1mm per 5mm camera travel.**

By analyzing both APE and RPE, we comprehensively evaluate the performance of the pose estimation process under various conditions, thereby understanding both the global positioning accuracy and the local consistency of motion tracking.

F. 3D Ultrasound Reconstruction

3D Ultrasound (US) image reconstruction has been used to validate the localization accuracy of the pose outputs from

previous steps in the pipeline. By stacking 2D US images in a scenario where the robot moves in a defined translational direction, we can effectively assess the pipeline's ability to track incremental positional changes.

This reconstruction method involves mapping image pixels to a 3D volume grid based on their corresponding pose data, emphasizing the translational aspects of movement. The process focuses on ensuring that the pixel intensities are accurately placed according to the spatial transformations dictated by the pose outputs.

III. Results

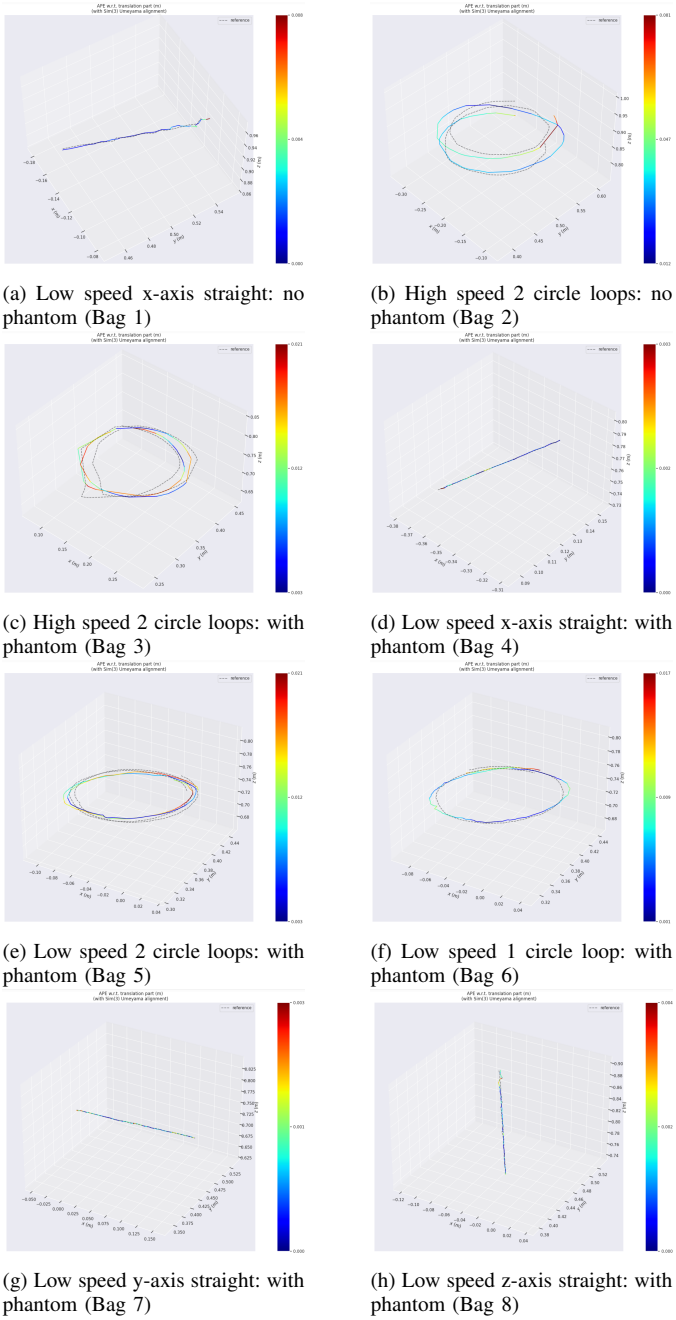
Eight distinct datasets have been collected to validate the accuracy of the SLAM outputs against the ground truth data using the EVO package. The datasets have been illustrated in Figure 4.

A. Analysis of Variable Impacts on ElasticFusion Performance

The datasets were strategically collected to systematically introduce variables such as speed, texture, trajectory shape, and the number of loops to ascertain their individual impact on the output. It has been observed that reducing the camera speed generally leads to improved results. For instance, there is a notable reduction in the Relative Pose Error (RPE) by approximately 6mm between Bag 3, which involves a faster robot speed, and Bag 5, characterized by a slower robot speed. This indicates that ElasticFusion results deteriorate with motion blur caused by high camera speed.

Furthermore, no significant impact on the RPE was observed in the comparisons of straight line trajectories (approximately equal to 4mm RPE) between Bag 1 (without phantom) and Bag 4 (with phantom), or in the circular trajectories (approximately equal to 12mm RPE) between Bag 2 (without phantom) and Bag 3 (with phantom). This shows that ElasticFusion performs well even in low texture scenarios which is expected as it is a direct vSLAM system.

Additionally, at lower robot speeds, the variables of trajectory shape and number of loops do not exhibit a discernible effect.



(a) Low speed x-axis straight: no phantom (Bag 1)

(b) High speed 2 circle loops: no phantom (Bag 2)

(c) High speed 2 circle loops: with phantom (Bag 3)

(d) Low speed x-axis straight: with phantom (Bag 4)

(e) Low speed 2 circle loops: with phantom (Bag 5)

(f) Low speed 1 circle loop: with phantom (Bag 6)

(g) Low speed y-axis straight: with phantom (Bag 7)

(h) Low speed z-axis straight: with phantom (Bag 8)

Fig. 4: Various datasets from experiments

B. ElasticFusion vs ORB-SLAM3 Performance

ORB-SLAM3 is a widely recognized simultaneous localization and mapping (SLAM) system, acclaimed for its versatility in handling monocular, stereo, and RGB-D cameras. It has established itself as a benchmark in SLAM technology, often serving as a reference point for comparative performance evaluations. In contrast, ElasticFusion is highlighted in the medical imaging literature [9] for its superior performance in low-texture environments. As a direct visual SLAM (vSLAM) system, ElasticFusion is particularly effective in scenarios where indirect vSLAM systems, such as ORB-SLAM3—which heavily rely on feature extraction and matching—may falter.

Illustrative results are shown in Figure 5. Observations indicate that the Relative Pose Error (RPE) per 5mm of trajectory for ElasticFusion ranges from 3.5 to 6 mm across all datasets, which is approximately four times better than that of ORB-SLAM3. This significant difference underscores ElasticFusion's robustness in handling scenarios with low-texture surfaces, such as those encountered with medical phantoms, aligning well with expectations based on its operational strengths.

C. Analysis after Sim(3) Alignment

The application of Sim(3) alignment is pivotal in enhancing the trajectory accuracy of both ElasticFusion and ORB-SLAM3 systems. Sim(3) adjustments compensate for scale, rotation, and translation discrepancies between the output

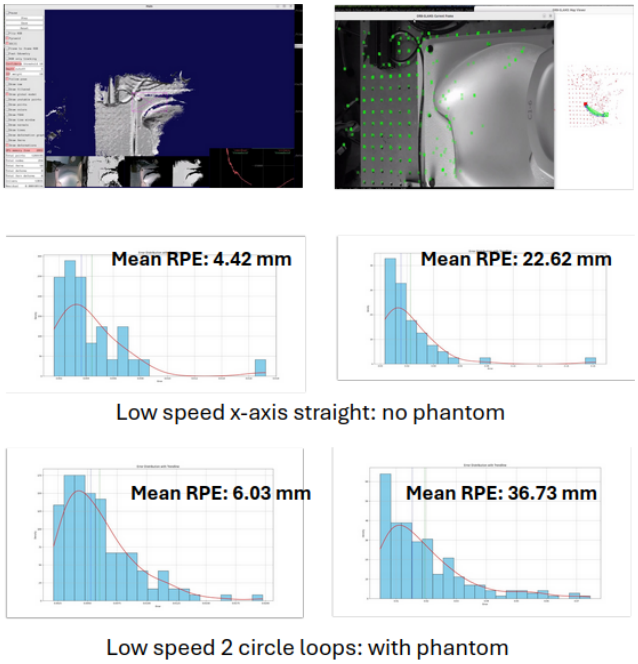


Fig. 5: ElasticFusion (left) vs ORB-SLAM3 (right)

trajectory and the ground truth. This precision is essential in environments where scale and orientation vary significantly.

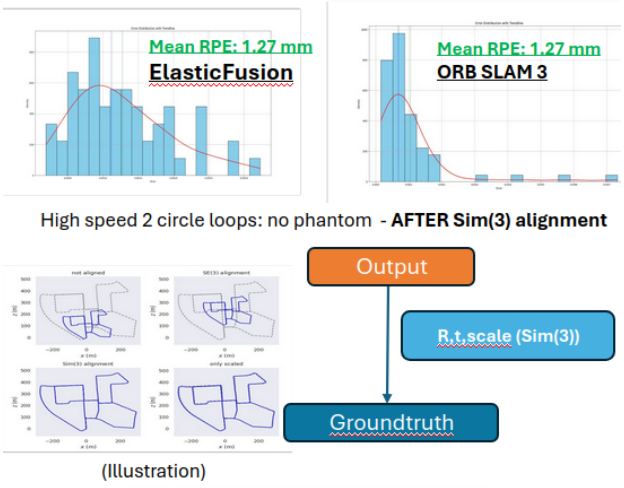


Fig. 6: Top: Relative Pose Error (RPE) distributions for ElasticFusion and ORB-SLAM3 following Sim(3) alignment for Bag 2, where both systems achieve a mean RPE of 1.27 mm, demonstrating highly accurate trajectory tracking. Bottom: Visual comparison of different alignment types.

As illustrated in Figure 6, the RPE distributions for both SLAM systems post-Sim(3) alignment are compared. Both systems consistently exhibit a mean RPE of 1.27 mm for Bag 2, underscoring the effectiveness of Sim(3) alignment in ensuring consistent and reliable trajectory accuracy across different SLAM systems. Across all datasets, Sim(3) alignment reduces the RPE to within 1-2 mm, indicating a significant improvement in SLAM performance.

Further analysis involved calculating the scale correction needed to align the trajectory outputs using Sim(3) alignment, found to be consistently close to 0.5 across all datasets. This uniformity suggests a stable scale adjustment for future datasets, as shown in Figure 7, or warrants an investigation into the reasons behind this consistency.

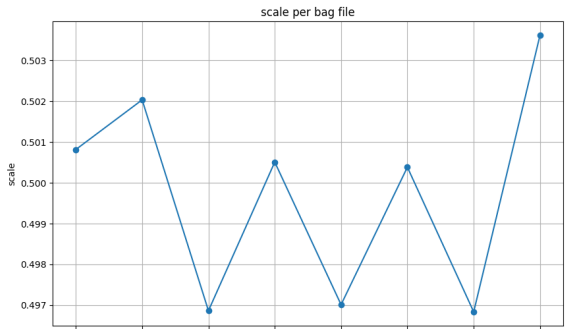


Fig. 7: Scale correction factors required for each dataset, from Bag 1 to Bag 8.

The necessity for ground truth data in Sim(3) alignment was also explored. Results indicate that utilizing only 50% of the ground truth data achieves satisfactory outcomes (mean APE of 3 mm). As detailed in Figure 8, increasing the proportion of ground truth data used progressively improves the APE, suggesting potential reductions in ground truth data requirements.

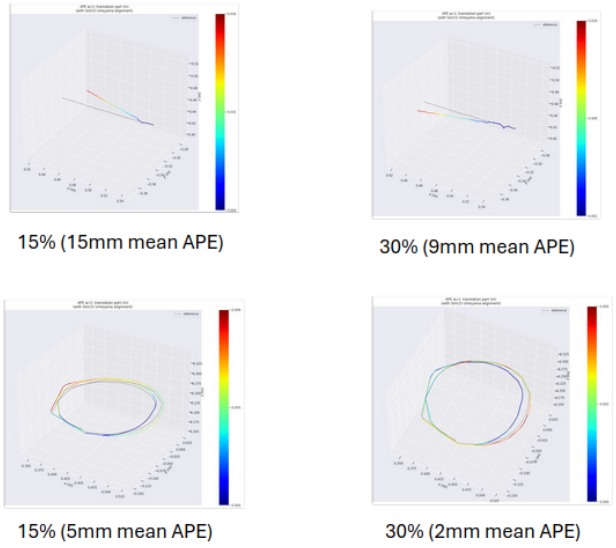


Fig. 8: Effect of increasing ground truth data usage on APE, demonstrating improvements from Bag 1 (Top) to Bag 5 (Bottom) after Sim(3) alignment.

D. Pose Graph Optimization (PGO)

As detailed in Section II D, the pose outputs of **Cam 2** have been utilized to refine the poses from **Cam 1** using Pose Graph Optimization. In this process, **Cam 1** poses are modeled as nodes, while the relative poses from **Cam 2** are represented as edges within the pose graph, employing the **g2o** optimization library.

Upon implementing PGO, an unexpected deterioration in accuracy was observed: the Mean Relative Pose Error (RPE) increased from 5.5 mm to 16.2 mm for a newly collected dataset involving a new circular trajectory, necessitated by the need to capture data from both cameras. This outcome is illustrated in Figure 9. The results suggest that integrating **Cam 2**'s relative poses as observations adversely affects the system's performance.

Although the initial results were not as anticipated, further experimentation is warranted. PGO remains a promising method for pose refinement, particularly due to its potential for enhancing system accuracy through the incorporation of a second camera.

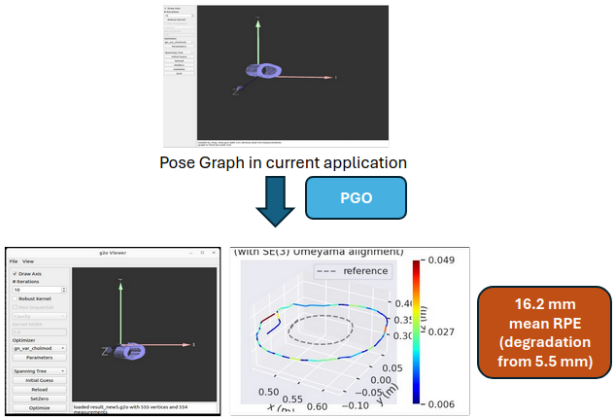


Fig. 9: (Top) Pose graph before PGO, (Bottom) Pose graph after PGO showing the resulting trajectory in EVO.

E. 3D US Reconstruction

As described in Section II F, 3D ultrasound (US) reconstruction was performed using a new straight-line trajectory facilitated by a custom fishing wire setup, depicted in Figure 10. This setup comprises a series of parallel wires intersected midway by two diagonal wires, all submerged in water to capture the US images. This configuration is particularly suitable for 3D US reconstruction due to the known distances between the wires, which serve as a benchmark for verifying the reconstruction accuracy.

The initial ultrasound image captured at the beginning of the run is shown in Figure 11.

Reconstruction results are presented in Figure 12. Both the outputs derived from SLAM and those from ground truth poses show considerable motion in the translation direction, as expected. This indicates a promising consistency in 3D US reconstruction when utilizing SLAM techniques. However,

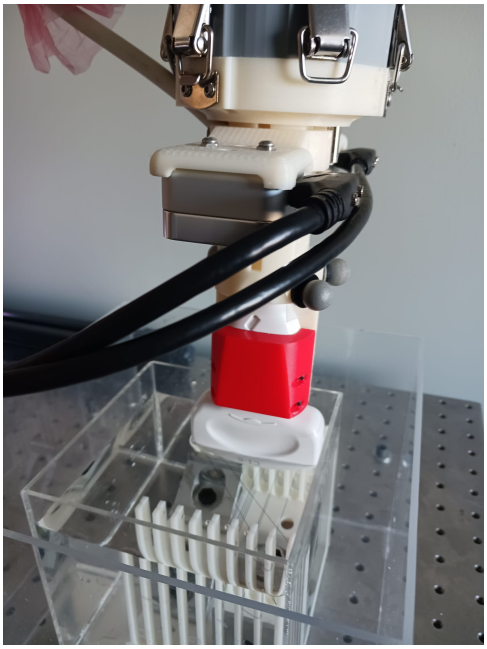


Fig. 10: 3D US fishing wire setup

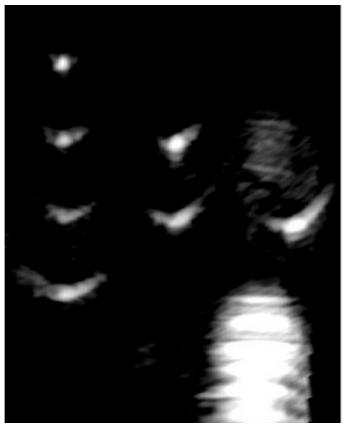


Fig. 11: 2D US image corresponding to the fishing wire setup

the integration of complete pose outputs for reconstruction remains the ultimate objective, which will further validate the efficacy of this approach.

The corresponding trajectory poses in EVO have been indicated for the ELasticFusion output and groundtruth in Figure 13. The noise in the ElasticFusion 3D reconstruction might be due to the slight deviation from the groundtruth indicated in the figure.

IV. Conclusion

This Directed Research project has successfully advanced the application of vSLAM technology in robotic 3D ultrasound imaging, demonstrating substantial improvements and efficiency. Below, are the principal achievements of this study:

1) **End-to-end Pipeline Development:** I successfully developed a robust end-to-end pipeline that transforms 2D ultrasound images and corresponding camera poses into

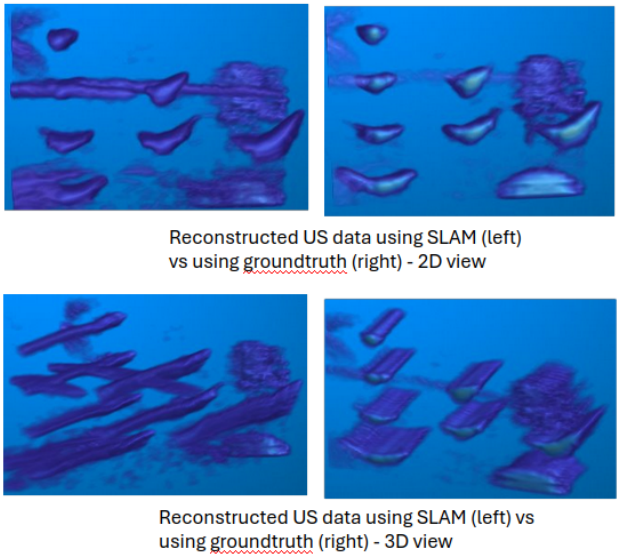


Fig. 12: 3D US reconstruction results from various perspectives (Top and Bottom views)

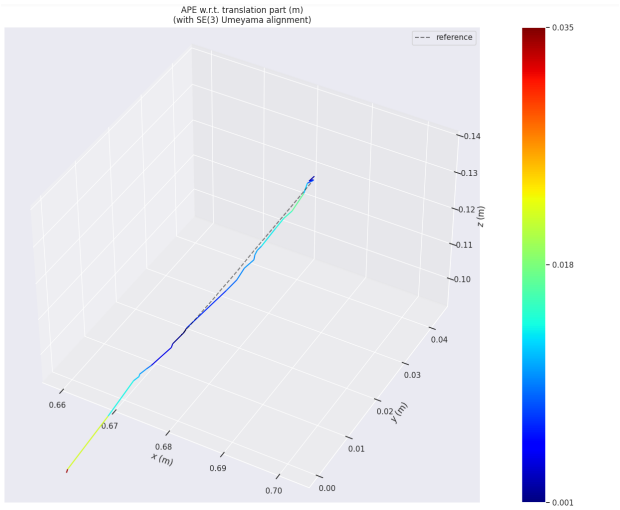


Fig. 13: ElasticFusion output (colored) vs groundtruth (dashed) poses.

detailed 3D ultrasound reconstructions. This pipeline enables a seamless transition from raw data to usable 3D models.

2) **Improvement in Localization Accuracy:** ElasticFusion demonstrated a significant enhancement in localization accuracy, with a Relative Pose Error (RPE) between 3.5 to 6 mm per 5mm trajectory, which is substantially better (four times lower) than the RPE observed with ORB-SLAM3.
3) **Refinement through Sim(3) Alignment:** The application of Sim(3) alignment led to an effective reduction in RPE to 1.27 mm, closely aligning with the target of achieving 1 mm RPE per 5mm travel. This demonstrates the benefit of consistent rotation, translation, and scale adjustments to initial pose outputs.
4) **Reduction in Groundtruth Data Requirements:** By employing only 50% of the groundtruth data for Sim(3) alignment, decent results have been obtained, with a mean Absolute Pose Error (APE) of 3 mm. This reduction in data dependency not only simplifies the operational demands but also enhances the system's efficiency.
5) **Demonstration of 3D Ultrasound Capabilities:** The results from 3D reconstruction experiments using 2D ultrasound images and output SLAM poses showcased the potential of ElasticFusion in clinical and diagnostic settings. These findings underscore the feasibility of using ElasticFusion for advanced 3D ultrasound applications, highlighting its adaptability and robustness.

In conclusion, the project's findings demonstrate that SLAM technologies, particularly ElasticFusion, hold significant promise for enhancing the precision and practicality of 3D ultrasound systems. The advancements in localization accuracy, combined with the ability to efficiently process ultrasound data into 3D reconstructions, pave the way for future innovations in medical imaging technology.

REFERENCES

[1] Mohamed F, Siang CV. A survey on 3D ultrasound reconstruction techniques. Artificial intelligence—applications in medicine and biology. 2019 Apr 27:73-92.
[2] Ito S, Ito K, Aoki T, Ohmiya J, Kondo S. Probe localization using structure from motion for 3D ultrasound image reconstruction. In2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017) 2017 Apr 18 (pp. 68-71). IEEE.
[3] M. Hastenteufel, M. Vetter, H.-P. Meinzer, and I. Wolf, "Effect of 3D ultrasound probes on the accuracy of electromagnetic tracking systems," Ultrasound in Med. Biol., vol. 32, no. 9, pp. 1359–1368, Sept. 2006.
[4] A.M. Goldsmith, P.C. Pedersen, and T.L. Szabo, "An inertial- optical tracking system for portable, quantitative, 3D ultra- sound," IEEE Int'l Ultrasonics Symp. Proc., pp. 45–49, Nov. 2008.
[5] S.-Y. Sun, M. Gilbertson, and B.W. Anthony, "Probe local- ization for freehand 3D ultrasound by tracking skin features," LNCS 8674 (MICCAI 2014), pp. 365–372, Sept. 2014.
[6] Ito S, Ito K, Aoki T, Ohmiya J, Kondo S. Probe localization using structure from motion for 3D ultrasound image reconstruction. In2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017) 2017 Apr 18 (pp. 68-71). IEEE.
[7] Zhang S, Zheng L, Tao W. Survey and evaluation of RGB-D SLAM. IEEE Access. 2021 Jan 21;9:21367-87.
[8] Qin X, Wang B, Boegner D, Gaitan B, Zheng Y, Du X, Chen Y. Indoor localization of hand-held OCT probe using visual odometry and real-time segmentation using deep learning. IEEE Transactions on Biomedical Engineering. 2021 Sep 29;69(4):1378-85.
[9] Turan M, Almalioglu Y, Araujo H, Konukoglu E, Sitti M. A non-rigid map fusion-based direct SLAM method for endoscopic capsule robots. International journal of intelligent robotics and applications. 2017 Dec;1:399-409.
[10] Whelan, T., Leutenegger, S., Salas-Moreno, R.F., Glocker, B., Davison, A.J.: Elasticfusion: Dense slam without a pose graph. In: Robot- ics: Science and Systems, Vol. 11 (2015)
[11] Newcombe, R.A, Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: real-time dense surface mapping and tracking. In: Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on, IEEE, pp. 127–136 (2011)