A Tracking and Mapping Method for Visually-degraded Environment

Yuxi Guo[1], Zhiqiang Dai[1], Xiangwei Zhu[1]

[1] School of Electronic and Communication Engineering, Sun Yat-sen University, Shenzhen, China
guoyx58@mail2.sysu.edu.cn, (daizhiqiang, zhuxw666)@mail.sysu.edu.cn

Keywords: Degraded Scene, Simultaneous Localization and Mapping, SLAM, Robustness, Three-dimensional Displays.

Abstract

When exploring uncertain areas, the system usually relies on Simultaneous Localization and Mapping (SLAM) to map the surrounding environment and track the location through the environment. Recent work has shown that the SLAM algorithm can exhibit strong robustness when observing distinctive features. However, some environments, such as tunnels or empty rooms, may lack sufficient features to navigate reliably. This is called environmental degradation, which refers to the difficulties encountered in positioning and mapping in specific scenarios and then lead to poor system performance. We propose an improved algorithm based on Non-Rigid Structure-from-Motion (NRSfM) and Deformable SLAM (DefSLAM). The method proposed in this paper can be used to deal with the degraded environment of visual sparse and repetitive features. In the experiment, the algorithm processes the close-up sequence of the degraded scene, both in the laboratory-controlled experiment and in the real-world sequence, to generate a 3D model of the scene relative to the mobile camera. It has been proved that the algorithm proposed in this paper achieves impressive performance on the dataset and the estimation method is robust to degradation.

1. Introduction

Visual simultaneous localization and mapping (VSLAM) algorithms is to locate a visual sensor in an unknown map which is being estimated simultaneously. When putting a robot at an uncertain location in an unknown environment, slam offers a way to let the robot gradually build a complete map of the whole environment, which refers to a map of the obstacle-free travel to every corner of the space. While a wide range of sensors can be used, from lidar to vision, cameras are finally chosen in this paper due to their low cost and rich information content.

The SLAM system first read and preprocess the image information captured from the camera. When an image is received, it is first processed to extract key-points and the associated descriptions. However, in many cases, it doesnt have sufficient features which are distinct enough to support the localization. Since the problems are inevitably to solve in environments with scarcity of texture features for vision sensors, it is important to make sure that the estimation methods are robust to degeneracy. The degeneracy is often due to the reduction of constraints. For example, if the constrains are single ,such as in a narrow corridor or tunnel environment(Yang et al., 2022), it will be difficult to judge the location of the robot even with the human eyes.

The main contribution of this paper is to propose an improved algorithm based on Non-Rigid Structure-from-Motion (NRSfM) and Deformable Visual SLAM. The method proposed can be used to deal with the degraded environments of sparse visual and non-rigid scenes.It exhibits impressive performance on the dataset and the estimation method is proved to be robust to the degeneracy.

The rest of this paper is organized as follows: Section II discusses the related work on slam for degraded environments, while Section III describes the proposed localizability model and the 3D map rebuilding method in detail. The experimental results are presented in Section IV. Finally, the paper will be given in Section V.

2. Related Work

2.1 Simultaneous localization and mapping

Nowadays, with rapid development of science and technology, more and more cutting-edge technologies are needed to improve users experience in virtual reality, mobile robots, drones, and automatic driving. Simultaneous localization and mapping(SLAM) is one of these technologies.

SLAM has become a fundamental part for modern applications(Klein and Murray, 2007a). Various sensors can be used for SLAM such as laser-range finders in (Hess et al., 2016) and (Li J, 2016) and different kinds of cameras in (Klein and Murray, 2007b) including monocular, stereo and RGB-D cameras.Cameras provide rich information and are much cheaper at the same time. Monocular cameras offer sequence of images captured by a single camera while RGB-D camera offers RGB images and depth images. With the availability of RGB-D cameras as shown in(Wang et al., 2019) and(Meng et al., 2018), the depth of the scenes can be used for the system (Sturm et al., 2012).

The research of monocular rigid VSLAM has been relatively mature. The most advanced monocular rigid VSLAM method in this field at this stage, such as (Mur-Artal et al., 2015), provides accurate, robust and fast results in robot scenarios. Innmann et al. (Grasa et al., 2014) proposed the EKF-SLAM algorithm, which uses Extended Kalman Filter(EKF) to calculate Kalman gain to optimize the robot pose and the position of feature points in the environment. Klein et al. (Mahmoud et al., 2019) obtained a dense map based on (Mur-Artal et al., 2015). The common point of these methods is to assume that the deformation of the scene can be ignored. Therefore, a purely rigid SLAM system can be achieved by removing any deformed scene area from the map.

SLAM in the deformed scenes needs to rely on sensors that can provide depth information. As a pioneering work of deformable VSLAM, DynamicFusion integrates frame-by-frame

depth information into a canonical shape, which incrementally maps the entire scene after partially observed exploratory trajectories. This standard shape is deformed to the current keyframe with the rigidest possible deformation model (Sorkine and Alexa, 2007). In (Innmann et al., 2016), the quality of deformation was improved by adding photometric error in the optimization.In (Song et al., 2018), the system optimizes the rigid system ORBSLAM (Mur-Artal et al., 2015) to achieve better trajectory and more robust deformable SLAM for medical endoscopic exploration.

The DefSLAM proposed recently (Lamarca, 2021) is the very first real-time monocular SLAM system that can be used under the deformation scenes. For the deformed scenes, the constraints are reduced, resulting in the environmental degradation.DefSLAM includes three parts mainly: the map, the deformation tracking thread and the deformation mapping thread. DefSLAM is able to generate 3D models of the corresponding scenes according to the information received from the camera sensors. It has demonstrated outstanding performance in achieving better feature extraction and matching ability than the traditional under degraded environment on Mandala dataset.

2.2 Environmental Degradation

Localization and mapping can usually be conducted accurately by the estimation methods. However, reliability of these methods is largely based on avoiding degeneracy that may arise from cases such as scarcity of texture features for vision sensors and lack of geometrical structures for range sensors.

The scene degradation refers to the cases when the performance of the system degrades due to the changes in the environment or sensor conditions. This may result from the presence of repetitive structures in the environment, increased noises or factors such as changes in the motion patterns.

Recently there are mainly two ways to solve the problems of environmental degradation. Change the methods of state estimation or add some state constraints to the state estimation methods. However, in the real application process, these two methods are not often completely satisfactory. On the one hand, alternative methods may not easy to find or not always exist. On the other hand, suitable state constraints can be hard to find, too.

Moreover, when the problem itself is not complicated and the problem is solvable, the increase of additional conditions will also bring about the increase of computational cost and finally lead to the unnecessary errors.

2.3 Non-Rigid Structure-from-Motion

Non-Rigid Structure from Motion (NRSfM) in (Parashar et al., 2018)is a computer vision and 3D reconstruction technique that designed to solve the problem of recovering the three-dimensional structure of non-rigid objects or degraded scenes from a sequence of two-dimensional images or video frames. It has good performance for dealing with the deformation of objects, which is inevitable to avoid in the degraded scenes.By parameterizing the model, NRSfM can effectively capture the motion and describe it accurately at the same time. NRSfM provides strong support for these occasions.

Compared with the traditional Structure from Motion (SfM) method, NRSfM has an outstanding performance in capturing the deformable nature of objects under degraded scenes. NRSfM uses techniques such as factorization, shape priors, temporal coherence and spatiotemporal regularization to build up an accurate 3D model which represents the dynamic behavior of the objects captured in the input images.

Shape-From-Template (SfT) methods recover the deformed object from monocular images and the objects textured 3D shape at rest. This textured shape-at-rest is called template. The method recovers the deformed shape by associatinf a deformation model with the corresponding template.The isometry assumption, first proposed in SfT methods, has also shown excellent results in NRSfM.

3. Proposed Method

3.1 System Overview

The system proposed can be roughly divided into three parts: the front-end, the back-end and the loop closure detection. The front-end of the system mainly includes visual odometry, which can estimate the camera poses at frame rate. It recovers the map points by minimizing the combination of reprojection error and deformation energy of each frame for tracking threads. The back-end of the system includes the nonlinear optimization part. When exploring the unknown regions, an extended mapping is performed to process keyframe for mapping. The loop closure detection is used to determine whether the robot has reached the previous position before. If it does, information will be provided to the back-end for processing.

By introducing a more advanced luminosity error model and developing a non-rigid deformation model of NRSfM, the ability of the proposed system to deal with the eigenvalues of the datasets improved greatly. The newly proposed algorithm has demonstrated its excellent ability to improve feature extraction and scenes reconstruction accuracy of visual SLAM in weak texture scenes on empty room datasets and no-texture sequences.

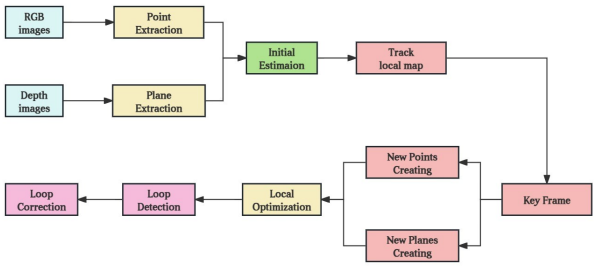

Figure 1. The overview of proposed system

The RGB images and depth images can be the inputs of the whole system. The front-end tracks camera pose using matched points and their descriptors. The back-end constructs and updates the maps of keyframes. Finally, the loop closure detection is used to determine whether the robot has reached the position before.

3.2 The Front-end

The front-end of the system performs a local odometry, matching the projection key points with the projection key points of the previous images. It estimates the transformation between the images. The common slam system is mostly based on the

set of points, using points to describe the scene to estimate the camera poses, usually choosing ORB feature points.

However, in most practical cases, for weak texture or even no texture scenes, the selection of ORB feature points can be difficult, which will lead to the decline of system performance. In addition, how to ensure the accuracy of data association is also a challenge for the system to obtain reliable results. The results of the system depend on the performance of the feature point detection and matching. Since the error of point measuring noise and the data association is cumulative, the performance of the system which relies only on the point feature in degraded environment is not good.

3.2.1 Data Association Visual odometry can be divided into feature point method and direct method according to whether it needs to extract features. The feature point method mainly includes feature extraction and matching as well as camera poses computing. The most commonly used feature extraction methods include SIFT algorithm, ORB algorithm and SURF algorithm.

The ORB algorithm is divided into two parts, namely feature point extraction and feature point description. The feature point extraction is developed from the Features from Accelerated Segment Test (FAST) algorithm and the feature point description is improved according to the Binary Robust Independent Elementary Features (BRIEF) feature description algorithm.FAST feature extraction believes that if a pixel has a big difference from the pixel of its neighborhood, it may be a corner point. BRIEF is a binary descriptor, which has a N-dimensional description vector consisting of 0 and 1. The vector describes the relationship between the gray values of each pixel pair randomly selected around the key points. The ORB feature combines the FAST feature point detection method with the BRIEF feature descriptor and improves them on the basis of the original, taking into account both the accuracy and calculation time.

In order to improve the performance of the system, we optimize the original DefSLAM algorithm in the selection of feature points. Since many weak texture scenes dont have enough and obvious corners but have large planes, the plane features and the corresponding descriptions are added while selecting ORB feature points to improve the performance of the system.

In the face of degraded scenes and structures, they often have obvious line and plane features, so this choice can help to achieve stable correlation and reduce cumulative errors at the same time.

Point-Plane SLAM Using Impressed Planes for Indoor Environments (Zhang et al., 2019)shows that the three-dimensional point cloud needs to be recovered from the depth map first. Then plane segmentation is performed on the point cloud and a small plane will be estimated at each point to determine the difference of each small plane. Finally, the plane parameters of the region are calculated, the parameters can be used directly to complete the data association when matching the plane feature. Planes with a cross relationship is generally a vertical plane or a matching surface.

The plane feature can be obtained from the depth map. It can be represented by a vector Π,

$$\Pi = (\mathbf{n}^T, d)^T \tag{1}$$

where $\mathbf{n}$ is the normal vector and d is the distance from the origin.

$$\mathbf{n} = (n_x, n_y, n_z) \tag{2}$$

For a point $\mathbf{P}$ on the plane π, we have

$$\mathbf{n}^T\mathbf{p} + d = 0 \tag{3}$$

Such a representation method is convenient for coordinate transformation,a plane in the world frame can be transformed into the camera frame.

Figure 2. The rgb image from the freiburg3_structure_notexture_near sequence of TUM Dataset

Figure 3. The corresponding depth image from the freiburg3_structure_notexture_near sequence of TUM Dataset

3.2.2 Camera Poses The front-end for data association is the core of the proposed algorithm. The system calculates the motion of the camera according to the information read from the images of adjacent time and then constructs the local map. It determines the poses of the current frame based on the observation of two adjacent frames and the pose of the previous frame.

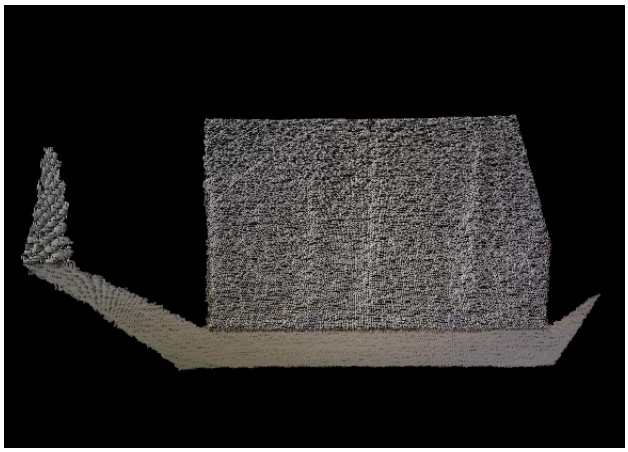

Figure 4. The corresponding point cloud from the freiburg3_structure_notexture_near sequence of TUM Dataset

Differ from the original technology, this paper proposes an improved Lucas-Kanade tracker.The traditional Lucas-Kanade algorithm estimates the optical flow from a series of images by tracking the motion of the object. The pyramid structure of the Lucas-Kanade feature tracker is shown in Figure 5. We assume that there is a variable size window in the system. When the image size is small, the window appears larger, and the optical flow can track the faster target.In the original image, the optical flow window is relatively small so the obtained optical flow is more accurate. By introducing a new iteration scheme, the solution can be refined at the rate of key frames.

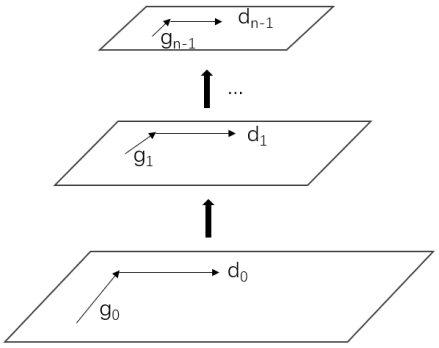

Figure 5. Structure of the Lucas-Kanade tracker

g_n and d_n are the remaining and initial optical flow direction at each layer respectively.

3.3 The Back-end

The back-end of the algorithm receives the camera poses from the visual odometry and the information of the loop closure detection.The image information will generate noise and errors during the matching process so it is necessary to optimize them for globally consistent trajectories and maps. At present, the mainstream optimization method is nonlinear optimization.

The map will be recovered to a surface. S_k is the plane at the reference frame k, which contains the map points observed in the keyframes before.It helps refine and create new points in the new keyframes.

3.3.1 Non-Rigid Structure from Motion(NRSfM) Isometric NRSfM assumes that each point of the plane will be modeled and any surface can be approximated to a small plane and maintained its curvature. NRSfM uses the warp kk between the keyframes k and k to solve the problem. The warp η_{kk} is a defined function that transforms a point in the anchor keyframe into the corresponding point in its covisible k, representing the transformation between two images.

$$\eta_{kk} : [x, y] \in R^2 \rightarrow [x^*, y^*] \in R^3 \tag{4}$$

The NRSfM is used to estimate a plane on a scale. We need to recover the scale and the scaled surface with respect to the previous map. Use the new surface and create a new template by calculating the triangular mesh and embedding the map points into the facets.

3.3.2 Extened Mapping Once the surface S_k is calculated, the key frame k is set to the reference keyframe. The shape observed in the current frame is different from the shape of the new template. If a new plane is observed, the system will create a new landmark.The system locally optimizes the camera poses and landmarks on the local map.After local optimization, redundant keyframes and planes with large errors will be deleted.

4. Experiments and Results

The method is implemented in C++. Prerequisites include Pangolin, OpenCV, Ceres library, PCL, DBoW2 and g2o. We use OpenCV to manipulate images and features. Pangolin is used for visualization and user interface. Ceres is for optimizing warp, running the NRSfM. The DBoW2 library is used to perform place recognition while G2o library is for non-linear optimizations. Unlike the traditional methods, the algorithm proposed in this article is optimized only in the mapped region so it can run on the CPU.

4.1 Datasets

4.1.1 TUM Dataset We choose the TUM dataset (Kuschk et al., 2017) (Cremers et al., 2017) to evaluate the performance of the SLAM systems.The TUM dataset contains a variety of scenes.They have different frame sizes and the camera move fast or slow. The dataset has objects with different structures and textures.The RGB image is 8-bit RGB images of 640x480 and the depth image is 16-bit images of 640x480 in PNG format.The freiburg3_structure_notexture_near sequence of TUM Dataset is chosen for the experiments.The pictures were taken in half a meter height along a zig-zag structure built from wooden panels. The object is fully wrapped in a white plastic foil with little to no texture.

4.1.2 Mandala Dataset The Mandala dataset is also used to evaluate the map quality of the proposed SLAM system in the degraded environment.It consists of five sequences(640Œ480 pixels), which contains five sequences.The degradation becomes more challenging with the number of dataset.

4.2 Results

The proposed algorithm is able to be applied to different weak texture degradation scenarios. The green points in 2D images represent the matching points while the black and red points in 3D images represent the corresponding matching points and the moving points respectively. Camera is located and shown as the green structure.

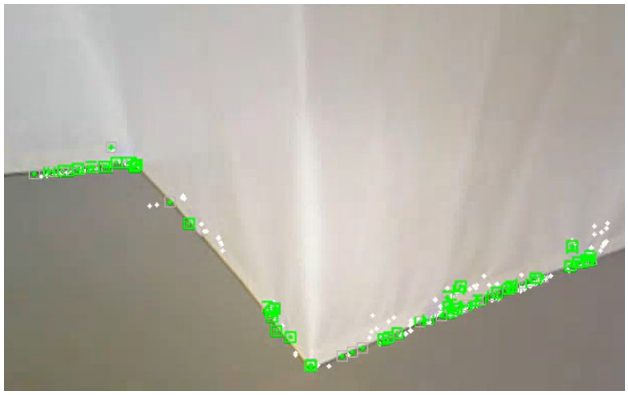

Figure 6. Performance on TUM dataset.The 2D images.

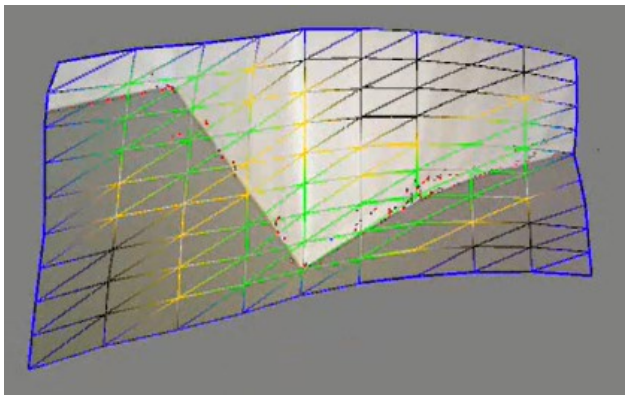

Figure 7. Performance on TUM dataset.The corresponding map.

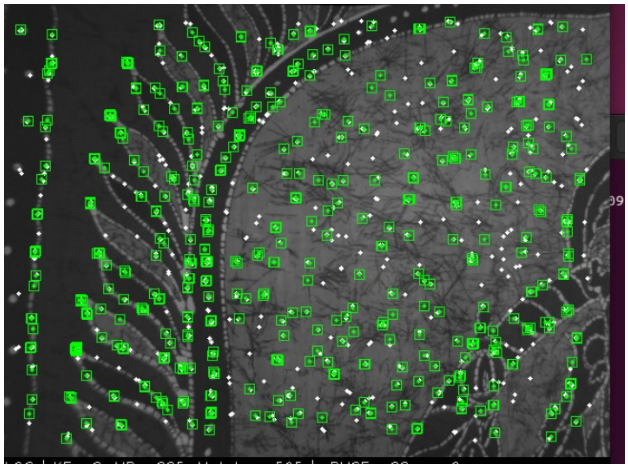

Figure 8. Performance on Mandala Dataset.The 2D images.

Our algorithm can work stably in degraded scenes with low contrast and highly similar features to provide clearer and more accurate images for scenes with weak texture degradation. The algorithm can not only improve the motion tracking logic of the camera in the weak texture degradation scenes, reducing the difficulty of data processing, but also improve the accuracy and quality of the pictures generated.

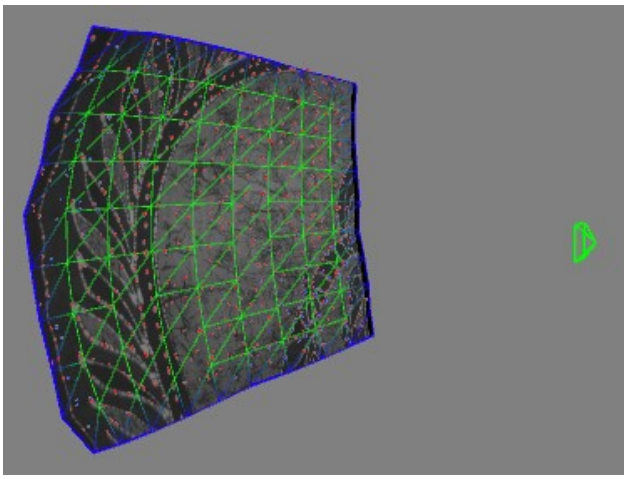

Figure 9. Performance on Mandala Dataset.The corresponding map.

The method proposed enhances the processing ability of weak texture scenarios, and improves the robustness and accuracy of the system in degraded environments.

5. Conclusion

In this paper, we propose an improved DefSLAM algorithm based on NRSfM and DefSLAM, which can be used to deal with visually sparse degraded environments.Experiments show that the algorithm proposed achieves satisfactory performance on the dataset and the estimation method has anti-degradation robustness.

The front-end of the algorithm mainly estimates the camera poses and scene degradation.Usually the point features are more often to use in Visual SLAM. However, only using point features can result in the lack of the robustness in the scenes such as low texture and structured environments. To overcome the above limitations, plane features are also employed.By adding the extraction of plane features, the system is more suitable for degraded environments. And after receiving the information of camera poses and loop closure detection from the front-end, the trajectory and map are obtained by optimizing the noise and error generated in the matching process. The results show that it is helpful for improving the robustness of SLAM system. However, it can not be directly compared with other systems, we focus on the description and introduction of the system itself.

The future work is to apply the system to medical endoscopic images.There are many challenges that are difficult to solve now in the field of medical images, such as the changing illumination, the organ deformation and usually poor texture. The light source is usually attached to the endoscope tip, which produces significant illumination variability, in addition to specular reflection in endoscopy. In general, surgical scenes are challenging for vision based reconstruction techniques and more future work need to be done.

References

Cremers, D., Leal-Taixé, L., Vidal, R., 2017. Deep Learning for Computer Vision (Dagstuhl Seminar 17391). *Dagstuhl Reports*, 7(9), 109–125.

Grasa, . G., Bernal, E., Casado, S., Gil, I., Montiel, J. M. M., 2014. Visual SLAM for Handheld Monocular Endoscope. *IEEE Transactions on Medical Imaging*, 33(1), 135-146.

Hess, W., Kohler, D., Rapp, H., Andor, D., 2016. Real-time loop closure in 2d lidar slam. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 1271–1278.

Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M., 2016. Volumedeform: Real-time volumetric non-rigid reconstruction. B. Leibe, J. Matas, N. Sebe, M. Welling (eds), *Computer Vision – ECCV 2016*, Springer International Publishing, Cham, 362–379.

Klein, G., Murray, D., 2007a. Parallel tracking and mapping for small ar workspaces. *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, 225–234.

Klein, G., Murray, D., 2007b. Parallel tracking and mapping for small ar workspaces. *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, 225–234.

Kuschk, G., d’Angelo, P., Gaudrie, D., Reinartz, P., Cremers, D., 2017. Spatially Regularized Fusion of Multiresolution Digital Surface Models. *IEEE Trans. Geosci. Remote. Sens.*, 55(3), 1477–1488.

Lamarca, J., P. S. B. A. . M. J. M. M., 2021. DefSLAM: Tracking and Mapping of Deforming Scenes From Monocular Sequences. *IEEE Transactions on Robotics*, 37(1), 291303.

Li J, Zhong R, H. Q. A. M., 2016. Feature-Based Laser Scan Matching and Its Application for Indoor Mapping. *Sensors (Basel)*, 16(8), 1265.

Mahmoud, N., Collins, T., Hostettler, A., Soler, L., Doignon, C., Montiel, J. M. M., 2019. Live Tracking and Dense Reconstruction for Handheld Monocular Endoscopy. *IEEE Transactions on Medical Imaging*, 38(1), 79-89.

Meng, X., Gao, W., Hu, Z., 2018. Dense RGB-D SLAM with Multiple Cameras. *Sensors*, 18(7). https://www.mdpi.com/1424-8220/18/7/2118.

Mur-Artal, R., Montiel, J. M. M., Tardós, J. D., 2015. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5), 1147-1163.

Parashar, S., Pizarro, D., Bartoli, A., 2018. Isometric Non-Rigid Shape-from-Motion with Riemannian Geometry Solved in Linear Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10), 2442-2454.

Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G., 2018. MIS-SLAM: Real-Time Large-Scale Dense Deformable SLAM System in Minimal Invasive Surgery Based on Heterogeneous Computing. *IEEE Robotics and Automation Letters*, 3(4), 4068-4075.

Sorkine, O., Alexa, M., 2007. As-Rigid-As-Possible Surface Modeling. *Eurographics*, 4, 109–116.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D., 2012. A benchmark for the evaluation of rgb-d slam systems. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 573–580.

Wang, R., Wan, W., Wang, Y., Di, K., 2019. A New RGB-D SLAM Method with Moving Object Detection for Dynamic Indoor Scenes. *Remote Sensing*, 11(10). https://www.mdpi.com/2072-4292/11/10/1143.

Yang, C., Chai, Z., Yang, X., Zhuang, H., Yang, M., 2022. Recognition of degradation scenarios for lidar slam applications. *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 1726–1731.

Zhang, X., Wang, W., Qi, X., Liao, Z., Wei, R., 2019. Point-Plane SLAM Using Supposed Planes for Indoor Environments. *Sensors*, 19(17). https://www.mdpi.com/1424-8220/19/17/3795.