Benchmark of Visual SLAM Algorithms: ORB-SLAM2 vs RTAB-Map

Nicolas Ragot, Redouane Khemmar, Adithya Pokala, Romain Rossi, Jean-Yves Ertaud

Benchmark of Visual SLAM Algorithms: ORB-SLAM2 vs RTAB-Map*

*

1st Nicolas Ragot
UNIRouen, Normandy University.
ESIGELEC/IRSEEM
Saint Etienne du Rouvray, France
nicolas.ragot@esigelec.fr

2nd Redouane Khemmar
UNIRouen, Normandy University
ESIGELEC/IRSEEM
Saint Etienne du Rouvray, France
redouane.khemmar@esigelec.fr

3rd Adithya Pokala
UNIRouen, Normandy University
ESIGELEC/IRSEEM
Saint Etienne du Rouvray, France
adithya.pokala@groupe-esigelec.org

4th Romain Rossi
UNIRouen, Normandy University
ESIGELEC/IRSEEM
Saint Etienne du Rouvray, France
romain.rossi@esigelec.fr

5th Jean-Yves Ertaud
UNIRouen, Normandy University
ESIGELEC/IRSEEM
Saint Etienne du Rouvray, France
jean-yves.ertaud@esigelec.fr

Abstract—This works deals with a benchmark of two well-known visual Simultaneous Localization and Mapping (vSLAM) algorithms: ORB-SLAM2 proposed by Mur-Atal & al in 2015 [7] and RTAB-Map proposed by [8]. The benchmark has been carried out with an Intel real-sense camera 435D mounted on top of a robotics electrical powered wheelchair running a ROS platform. The ORB SLAM has been implemented taking into account a monocular, stereo and RGB-D camera. RTAB SLAM, meanwhile, has only implemented with monocular and RGB-D camera. Several experiments have been carried out in a controlled indoor environment at the ESIGELEC's Autonomous Navigation Laboratory. These experiments are supported by the use of the VICON motion capture system used as a ground-truth to validate our results [1]. Different motion scenarios are used to test and benchmark the SLAM algorithms in various configurations: straight-line, straight-line and back, circular path with loop closure, etc.

Index Terms—Visual SLAM, ORB SLAM, RTAB SLAM, Localization, Mapping, Visual Odometry.

I. INTRODUCTION

Estimating the camera trajectory while simultaneously reconstructing the environment is a key and well-know problem in robotics and computer vision. This can be achieved by the use of different measuring means (LIDAR, RADAR) and algorithms (particle filter, extended Kalman filter, Graph-SLAM). When cameras are used, SLAM techniques are named as visual SLAM (vSLAM). vSLAM algorithms have been widely proposed in the field of computer vision, robotics. Taketomi & al. [6] proposes an exhaustive survey of real-time

vSLAM algorithms developed from 2010 to 2016. In [2], a very exhaustive state of the art about the vSLAMis presented. Obtaining a complete vSLAM system essentially consists in adding the cartography or the geometric consistency of the visual odometry map. The pioneering approach of visual odometry was presented by Nister et al. [2] in 2003 for monocular and stereo cases. They presented estimation within RANSAC for the first time as well as the estimation from 2D-3D matches instead of triangulating for each stereo pair which allowed to have more robustness [11]. Indeed, the majority of the methods follow the indirect scheme where they carry out features detection and movement estimation within RANSAC. However, direct methods have also been implemented such as SVO [12] and DSO [13].

Location-based on existing maps could also be done without the on-line mapping process or building a map beforehand. In recent years, specific card formats have been proposed to meet the need for prior knowledge for the autonomous car such as OpenStreetMap (OSM). Guo et al. proposed in [14] a low-cost approach based on OSM maps, INS, GPS and spelling images generated from a camera. A second optimization using visual odometry is performed [15].

Location-based on recognition of places: usually done in two phases: an off-line learning phase where a geotagged database is built and the location is done online by finding the image that most closely resembles the current one. Site recognition is also an essential part of a visual SLAM system to establish loop closures on one hand and to relocate on the other hand in the event of loss of tracking. The visual points of interest are further subdivided into two categories: those that select parts of the image and those that use the entire image without resorting to a selection step. Examples of the first category are the SIFT [16] and SURF [17] descriptors, and

the second category global descriptors such as HOG [19] and GIST [18]. bag-of-words separates the point of interest space into a finite number of visual words. FabMap [20] [21] which is a probabilistic approach that has shown its effectiveness on a trajectory of 1000 km in length. Other work has been based on global descriptors such as HOG in [22] and GIST in [23]. Global descriptors are more dependent on the point of view than local descriptors, but they do not work well with changes in brightness. The most representative approach in this category is SeqSLAM [24] which has demonstrated a much better performance than FabMAP under conditions of extreme atmospheric changes and brightness. SeqSLAM is based on the use of subsampled whole images by searching in a database that is built in the sequence that most closely resembles the current sequence acquired. An extension of SeqSLAM for 2D interior environments has been proposed in [25].

We therefore present in this paper a benchmark comparative study between two visual SLAM approaches: ORB SLAM and RTAB SLAM. The ORB SLAM has been implemented taking into account different kinds of sensors such as monocular, stereo and RGB-D camera. The RTAB SLAM, meanwhile, it has been experimented taking into account both monocular and RGB-D images. The wheelchair has been modified to fit the developed platform. A new motor controller has been developed, the wheelchair is connected within a wireless communication system, and the instrumentation is carried out within a Realsense camera including a monocular, stereo and a RGB-D sensors. A ground truth data based VICON system is used to validate the results obtained.

This paper is organized as follows: Section 1 introduces the motivation of the paper. Section 2 presents the state of the art about SLAM. Section 3 presents the wheelchair architecture and how to control the wheelchair. Section 4 illustrates the comparison between ORB SLAM and RTAB SLAM within different kind of camera like monocular, stereo and RGB-D camera. In the same section, we present how we have validated the approach developed by comparison between SLAM localization and ground truth localization carried out by VICON system. The results obtained will be presented in Section 5. Finally, in Section 6, we will conclude this paper.

II. RELATED WORKS

The literature review suggests to classify the methods in two main categories: one is related to the *indirect methods* whereas the other deals with the *direct ones*. The indirect methods are based on the detection of points of interest in the image (features) and the establishment of 2D-3D correspondences between the points of interest and the map by matching/tracking. The pose of the camera is then obtained by solving the Perspective-n-Point (PnP) problem. It is noted that the majority of vSLAM methods assume that the camera is previously calibrated, which implies that the intrinsic parameters of the camera are previously known [3].

In robotic mapping and navigation, Simultaneous Localization and Mapping (SLAM) is the computational problem of constructing or updating a map of an unknown environment

while simultaneously keeping track of an agent's location within it. While this initially appears to be a chicken-and-egg problem there are several algorithms known for solving it, at least approximately, in tractable time for certain environments. Popular approximate solution methods include the particle filter, extended Kalman filter, and GraphSLAM.

Indirect methods-based feature detection belong to other subcategories of vSLAM namely filter-based SLAM and optimization-based SLAM. A pioneering method is MonoSLAM presented in 2007, allowing to simultaneously estimate the pose of the camera and the 3D structure of an unknown environment using an Extended Kalman filter (EKF) [4]. The 3D camera movement (6 DOF) and the 3D positions of the features are represented by a state vector of EKF. PTAM offers an improvement by separating tracking and mapping in different *threads* of the CPU [5]. PTAM also introduced the use of keyframes for mapping as well as a feature classifier to search for the keyframe closest to the input image. ORB-SLAM can be considered as an extension of PTAM including features such as separate *threads* and keyframes, the detection of closing loops and the optimization of the poses graph [7]. ORB-SLAM [7], introduced in 2015, is the most complete indirect monocular vSLAM method and has been extended for stereo systems and RGB-D systems in ORB-SLAM2 in 2017 [8].

Unlike feature-based indirect methods, use directly pixel intensity of input images. A completely direct method is DTAM which, like PTAM, has two distinct parts: pose tracking and mapping [9]. However, PTAM tracks a set of 3D points while DTAM maintains dense depth maps for a selection of keyframes. The most representative method in the category of direct methods is LSD-SLAM (Large-Scale Direct SLAM) which is a monocular SLAM algorithm [10]. The map is constructed based on keyframes composed of camera image, a map of the inverse depths (inverse depth map) and the variance of the inverse of depths.

III. ROBOTICS WHEELCHAIR ARCHITECTURE

The robotics electrical powered wheelchair from the ESIG-ELEC's lab is an Invacare, model Bora from which all the proprietary electronics has been removed and replaced by:

- Embedded PC running a Linux Ubuntu 16.04 LTS distribution
- Motor driver from Roboteq
- Xbox joystick in place of the original one which can be connected to the wheelchair through a USB or Bluetooth connection.
- WIFI router for getting a wireless local network on the wheelchair
- Embedded HMI with a touch-screen panel

The wheelchair software is fully developed under the ROS robotics middleware. Figure 1 shows the wheelchair description. An embedded computer is used as a heart of processor in the wheelchair. It possesses 8 GB of RAM and 250 GB of SSD. It runs Ubuntu 16.04 as operating system. A tactile screen is connected to the embedded computer using the
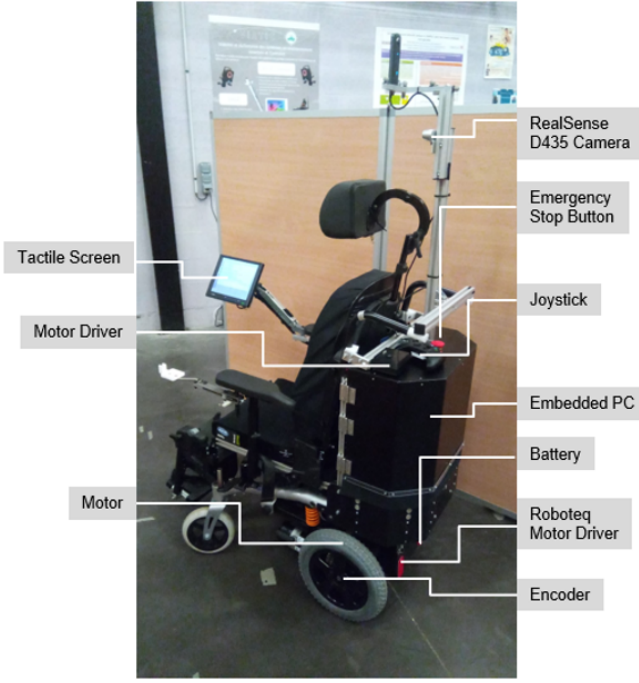
Fig. 1. ADAPT Real Wheelchair Platform.

HDMI cable to present the user interface to the user. A motor driver from Roboteq is used to drive motors. The wheelchair also has its own Wi-Fi router. An Xbox One joystick is used to control the wheelchair. A RealSense D435 depth camera from Intel is used for embedded vision. Figure 2 shows the hardware architecture of the wheelchair.
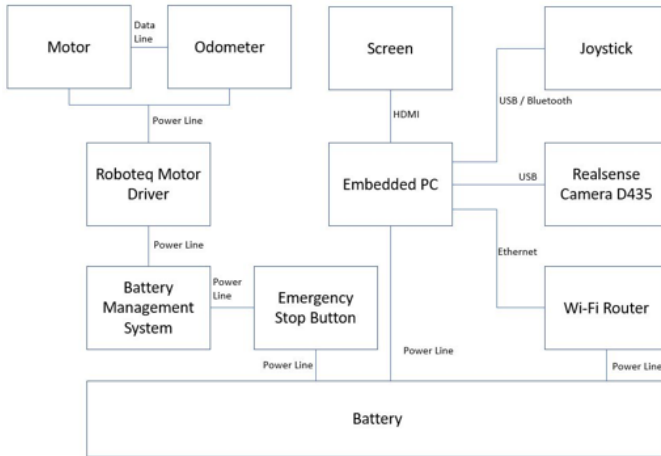

Fig. 2. Hardware Architecture of the Wheelchair.

IV. VSLAM COMPARISON BETWEEN ORB-SLAM AND RTABMAP SLAM

A. ORB-SLAM

ORB-SLAM2 is a real-time SLAM library for monocular, stereo and RGB-D cameras that computes the camera trajec-

tory and a sparse 3D reconstruction. It is able to detect loops and re-localize the camera in real time. The system works in real-time on standard CPUs in a wide variety of environments from small hand-held indoors sequences, to drones ying in industrial environments and cars driving around a city. The back-end based on bundle adjustment with monocular and stereo observations allows for accurate trajectory estimation with metric scale. The system includes a lightweight localization mode that leverages visual odometry tracks for unmapped regions and matches to map points that allow for zero-drift localization. The main functionalities of ORB SLAM are: feature tracking, mapping, loop closure and localization.

1) Monocular ORB-SLAM 2: Monocular SLAM requires just a single camera. The ORB SLAM converts the image to grayscale for its application. The first step is to detect features and initialize the map and its position. Once it gets initialized, it starts creating map. Monocular SLAM requires a procedure to create an initial map because depth cannot be recovered from a single image. One way to solve the problem is to initially track a known structure. In the context of ltering approaches, points can be initialized with high uncertainty in depth using an inverse depth parametrization, which hopefully will later converge to their real positions. The figure below shows the flow of process in ORB-SLAM monocular.

2) Stereo ORB-SLAM 2: Stereo ORB-SLAM requires a stereo camera. The feature detection in stereo SLAM is better than monocular SLAM. The initialization is fast and has many advantages over monocular SLAM. For stereo cameras, we extract ORB in both images and for every left ORB we search for a match in the right image. This can be done very efficiently assuming stereo rectified images, so that epipolar lines are horizontal. We then generate the stereo keypoint with the coordinates of the left ORB and the horizontal coordinate of the right match, which is subpixel refined by patch correlation.

3) RGB-Depth ORB SLAM 2: RGB-D SLAM requires an RGB image with its depth image. For RGB-D cameras, we extract ORB features on the RGB image, for each feature with coordinates $(uL; vL)$ we transform its depth value d into a virtual right coordinate:

$$uR = uL - \frac{fx * b}{d} \quad (1)$$

where fx is the horizontal focal length and b is the baseline between the structured light projector and the infrared camera. The uncertainty of the depth sensor is represented by the uncertainty of the virtual right coordinate. In this way, features from stereo and RGB-D input are handled equally by the rest of the system.

4) MAP Save and Load: A feature to save the map generated by the ORB SLAM is developed. Ability to load the map for localization purpose was also developed. All keyframes, features and its descriptors are saved to a binary file. After every mapping session the algorithm asks the user whether to save the map or not. At the beginning of SLAM, the algorithm asks the user whether to load the saved map to

begin localization or mapping within the environment which is already built.

B. RTABMAP SLAM

RTAB-MAP stands for Real Time Appearance Based Mapping. It is distributed as an open source library since 2013. RTAB-Map started as an appearance-based loop closure detection approach with memory management (shown in below figure) to deal with large-scale and long-term online operation. It then grew to implement Simultaneous Localization and Mapping (SLAM) on various robots and mobile platforms. RTAB-Map supports both visual and lidar SLAM, providing in one package a tool allowing users to implement and compare a variety of 3D and 2D solutions for a wide range of applications with dierent robots and sensors. It uses depth image with RGB images to construct maps. The graph is created here, where each node contains RBG and depth images with corresponding odometry pose. The links are transformation between each node. When the graph is updated, RTAB-Map compares the new image with all previous ones in the graph to find a loop closure. When a loop closure is found, graph optimization is done to correct the poses in the graph. For each node in the graph, we generate a point cloud from the RGB and depth images. This point cloud is transformed using the pose in the node. The 3D map is then created [11].

V. RESULTS ANALYSIS

The tests were carried out in LNA lab of IRSEEM for indoor scenarios and in car parking of IRSEEM for outdoor scenarios. LNA lab is equipped with Vicon camera system for tracking the wheelchair position and will be used as ground truth to evaluate the distance measurement of SLAM.

A. ORB SLAM

ORB SLAM works perfectly except with monocular camera as it loses points during mapping and also initializes very slowly. It can estimate its odometry and track features efficiently. The ORB SLAM creates a sparse 3D map as shown in the Figure 3.

The ORB SLAM creates a sparse 3D map. ORB SLAM is successful in creating a map and generating its trajectory in the environment. It can also detect closed loops successfully. The Figure 4 shows the trajectory created by ORB SLAM in indoor environments. Figure 5 shows the trajectory in outdoor environments.

The distance estimation is also accurate to the ground truth measurement by the Vicon. The only problem with ORB SLAM with monocular camera was that it was slow at initialization and also it loses points during the mapping. Due to this the ORB SLAM loses its position and will not be able to continue with mapping. The Figure 6 shows an example where a trajectory is generated by ORB SLAM with monocular in comparison with ORB SLAM with stereo and RGBD.

The Table I below shows the distance measurement for various paths. For the indoor scenarios we have used Vicon
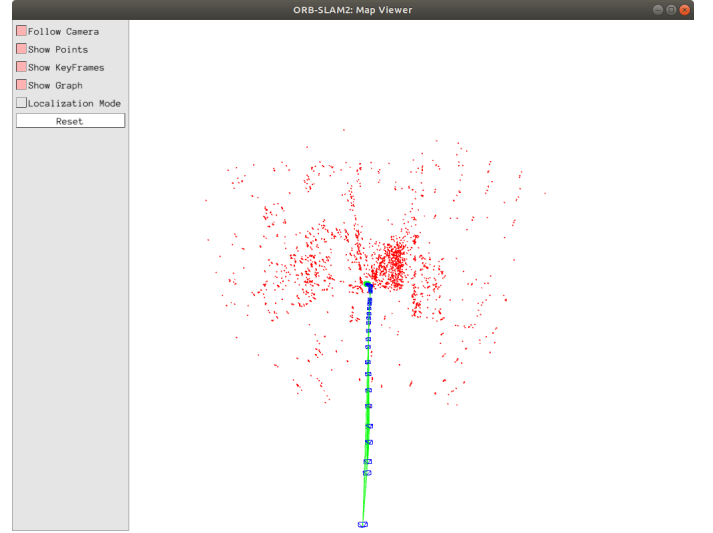

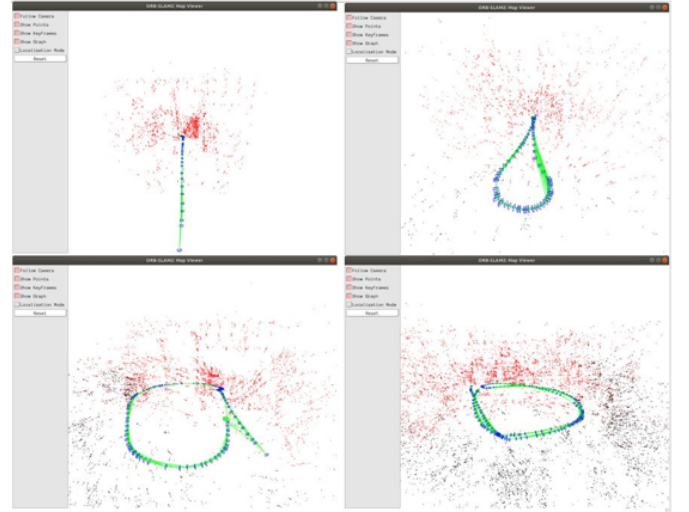Fig. 3. An example of a map created by ORB SLAM

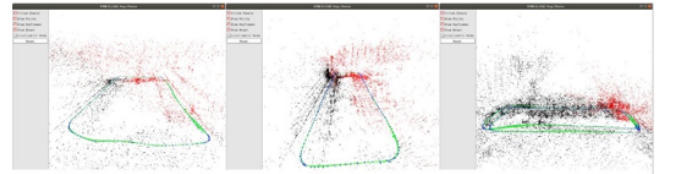Fig. 4. Map and Trajectory estimation of ORB SLAM in indoor environment

Fig. 5. Map and Trajectory estimation of ORB SLAM in outdoor environment

as the ground truth and for outdoor scenarios we have used measuring tape as the ground truth (approximate distance). The ORB SLAM with monocular camera had trouble initializing and also could not complete the trajectory. This is a limitation with monocular cameras and only alternative is to use either stereo camera or RGB-D camera. From Table I we can observe
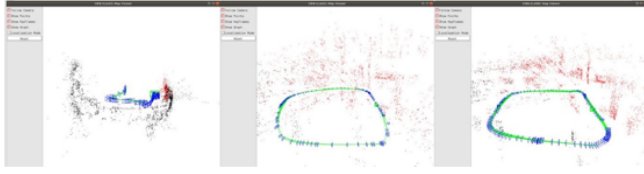
Fig. 6. ORB SLAM trajectory comparison between Monocular, Stereo and RGBD cameras.

TABLE I
ORB SLAM AND VICON DISTANCE (IN METERS) COMPARISON.

Scenario	Ground Truth	ORB Stereo	ORB RGBD
Indoor 1	7.82	7.69	7.16
Indoor 2	6.98	7.38	6.81
Indoor 3	15.05	15.49	14.32
Indoor 4	21.75	21.17	21.31
Indoor 5	29.69	21.47	22.13
Indoor 6	37.69	38.84	38.79
Outdoor 1	28	28.85	19.27
Outdoor 2	28	27.29	17.83
Outdoor 3	56	56.63	37.25
Outdoor 4	92	94.55	65.87
Outdoor 5	92	95.8	66.97
Outdoor 6	80	84.38	66.24
Outdoor 7	80	81.9	63.36
Outdoor 8	160	172.6	129.41

the results of ORB SLAM with stereo and RGBD. In indoor scenarios, ORB SLAM was accurate with both stereo and RGBD cameras but in outdoor scenarios, stereo camera has shown better results compared to the RGBD.

A comparison of trajectories in two indoor scenarios by ORB SLAM with ground truth trajectory by Vicon was done. The trajectories generated by ORB SLAM was closely accurate to that of the ground truth trajectory by Vicon. The Figure 7 shows the trajectory comparison.
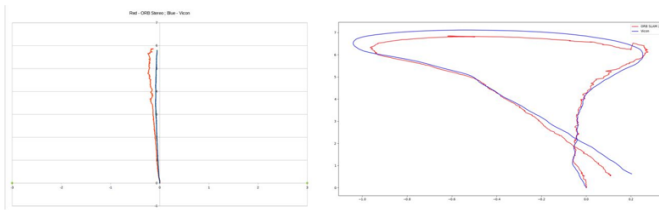

Fig. 7. ORB SLAM and Vicon trajectory comparison.

B. RTABMAP SLAM

Unlike ORB SLAM, RTABMAP SLAM creates a dense 3D map of the environment. It is tested on RGBD camera only. The figure below shows an example of the map created by the RTABMAP SLAM. RTABMAP was tested in indoor scenarios. It was successful in creating a map and detecting loops. The Figure 8 shows the map and trajectory estimation of RTABMAP.
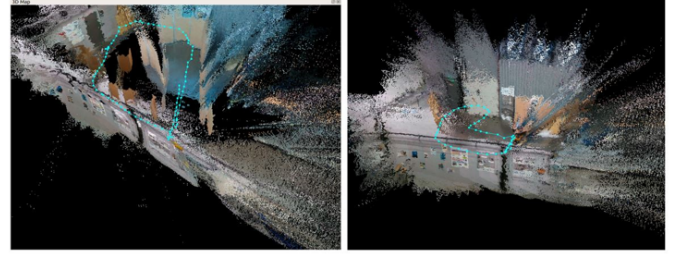

Fig. 8. Map and trajectory estimation by RTABMAP SLAM.

The Table II shows the RATBMAP SLAM measurement of distance in comparison with Vicon measurement. From above

TABLE II
RTABMAP DISTANCE COMPARISON WITH VICON.

Scenario	Ground Truth	RTABMAP RGBD
Indoor 1	6.48	11.83
Indoor 2	24.5	37.61
Indoor 3	23.26	31.31

results you can see that the mapping and trajectory estimation of RTABMAP is very good but the distance measurement is not. The distance estimated by the RTABMAP is not better than that of ORB SLAM.

C. ORB SLAM vs RTABMAP

The ORB SLAM was compared with RTABAMP in same scenarios as RTABMAP was tested. The Table III shows the distance measurement comparison between two SLAM algorithms. From the above table we can say that the distance measurements by ORB SLAM are more accurate than that of RTABMAP. From the above table we can say that the distance

TABLE III
ORB SLAM AND RTABMAP COMPARISON.

Scenario	Ground Truth	RTAB RGBD	ORBSLAM Stereo	ORBSLAM RGBD
Indoor 1	6.48	11.83	5.33	8.94
Indoor 2	24.5	37.61	30.64	27.78
Indoor 3	23.26	31.31	20.82	21.43

measurements by ORB SLAM are more accurate than that of RTABMAP.

VI. CONCLUSION

In this paper, we have presented a wheelchair localization and mapping-based comparison of two vSLAM algorithms, and within both of indoor and outdoor environment. Different approaches in VSLAM were developed around ORB SLAM and RTAB SLAM. In order to improve the quality of localization and mapping obtained by the basic SLAM, we have updated the ORB SLAM to our system by taking into account some parameters. We have tested ORB SLAM using Monocular camera, Stereo camera and RGBD camera.

we also added a feature where ORB SLAM prompts the user for saving the map and loading the map for pure localization. The ORB SLAM is able to save map and later load the map for localization purpose. The ORB SLAM has also been improved to estimate an accurate trajectory and also provide accurate odometry estimations. However, ORB SLAM with Monocular camera could not provide good trajectory or distance estimations due to the limitations of Monocular camera. The RTABMAP is also tested and used to compare with ORB SLAM. The trajectory estimation by RTABMAP is accurate but the odometry estimation is not as accurate as ORB SLAM. Few tests show that RTABMAP is better at trajectory estimation when compared to ORB SLAM but when it comes to distance measurement ORB SLAM is better. We have validated the whole developments under wheelchair platform within comparison between SLAM data, RTAB data, and ground truth data-based VICON obtained in indoor environment. Other scenario was carried out in outdoor environment for the validation of the platform.

References

[1] M. Merriaux, and Y. Dupuis, and R. Boutteau, and P. Vasseur, and X. Savatier, 'A Study of Vicon System Positioning Performance', Sensors Book, 2017.

[2] S. Ouerghi, Embedded Architecture for Vision-based Mobile Localization', PHD Thesis, Carthage University, SUP'COM: Ecole Suprieure des Communications de Tunis, 16 January, 2019.

[3] S. Ouerghi, Mobile Robot Localization State of the Art', Research Paper, ESIGELEC/IRSEEM, 24 January, 2019.

[4] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, MonoSLAM: Real-Time Single Camera SLAM, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 10521067, Jun. 2007.

[5] G. Klein and D. Murray, Parallel Tracking and Mapping on a Camera Phone, in Proc. Eigth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR09), Orlando, 2009.

[6] T. Taketomi, H. Uchiyama, and S. Ikeda, Visual SLAM algorithms: a survey from 2010 to 2016, IPSJ Transactions on Computer Vision and Applications, June. 2017.

[7] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, ORB-SLAM: A Versatile and Accurate Monocular SLAM System, IEEE Transactions on Robotics, vol. 31, no. 5, pp. 11471163, Oct. 2015.

[8] R. Mur-Artal and J.D. Tardos, 'Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras', IEEE Transactions on Robotics, vol. 33, no. 5, pp. 12551262, 2017.

[9] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, DTAM: Dense tracking and mapping in real-time, 2011, pp. 23202327.

[10] J. Engel, T. Schps, and D. Cremers, LSD-SLAM: Large-Scale Direct Monocular SLAM, in Computer Vision ECCV 2014, 2014, pp. 834849.

[11] D. Scaramuzza and F. Fraundorfer, Visual Odometry [Tutorial], IEEE Robotics & Automation Magazine, vol. 18, no. 4, pp. 8092, Dec. 2011.

[12] Forster C, Pizzoli M, Scaramuzza D (2014) SVO: fast semi-direct monocular visual odometry. In: Proceedings of International Conference on Robotics and Automation. pp 1522

[13] Engel J, Koltun V, Cremers D (2016) Direct sparse odometry. CoRR. abs/1607.02565

[14] C. Guo, J. i Meguro, Y. Kojima, and T. Naito, Automatic lane-level map generation for advanced driver assistance systems using low-cost sensors, in 2014 IEEE International Conference on Robotics and Automation (ICRA), 2014, pp. 39753982.

[15] C. Guo, K. Kidono, J. Meguro, Y. Kojima, M. Ogawa, and T. Naito, A Low-Cost Solution for Automatic Lane-Level Map Generation Using Conventional In-Car Sensors, IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 8, pp. 23552366, Aug. 2016.

[16] D. G. Lowe, Object recognition from local scale-invariant features, in Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, vol. 2, pp. 11501157 vol.2.

[17] H. Bay, T. Tuytelaars, and L. V. Gool, Surf: Speeded up robust features, in In ECCV, 2006, pp. 404417.

[18] A. Oliva and A. Torralba, Building the gist of a scene: the role of global image features in recognition, Prog. Brain Res., vol. 155, pp. 2336, 2006.

[19] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05), 2005, vol. 1, pp. 886893 vol. 1.

[20] M. Cummins and P. Newman, FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance, The International Journal of Robotics Research, vol. 27, no. 6, pp. 647665, Jun. 2008.

[21] M. Cummins and P. Newman, Appearance-only SLAM at large scale with FAB-MAP 2.0, The International Journal of Robotics Research, vol. 30, no. 9, pp. 11001123, Aug. 2011.

[22] C. McManus, B. Upcroft, and P. Newmann, Scene Signatures: Localised and Pointless Features for Localisation, 2014.

[23] C. Siagian and L. Itti, Biologically Inspired Mobile Robot Vision Localization, IEEE Transactions on Robotics, vol. 25, no. 4, pp. 861873, Aug. 2009.

[24] M. Milford and G. Wyeth, SeqSLAM: visual route-based navigation for sunny summer days and stormy winter nights, in Proceedings of the 2012 IEEE International Conferece on Robotics and Automation (ICRA), River Centre, Saint Paul, Minnesota, 2012, pp. 16431649.

[25] J. Mount and M. Milford, 2D visual place recognition for domestic service robots at night, in 2016 IEEE International Conference on Robotics and Automation (ICRA), 2016, pp. 48224829.

[26] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, Visual odometry and mapping for autonomous flight using an RGB-D camera, in Proc. of the Int. Sym. of Robot. Research, Flagstaff, AZ, Aug. 2011.