A Comparative Analysis of Visual Localization Algorithms in Indoor Navigation

İrem Yakar, Esra Gunaydın, Ramazan Alper Kucak, Serdar Bilgi, Mahmut Oguz Selbesoglu

ITU, Department of Geomatics Engineering, 80626 Maslak Istanbul, Türkiye - (yakari, esgunaydin, kucak15, bilgi, selbesoglu) @itu.edu.tr

Technical Commission II

Keywords: Visual Odometry, EKF SLAM, Indoor Navigation, Localization

Abstract

Localization can be defined as the process of determining the position and orientation of an entity within an environment, that would enable it to navigate and carry out tasks effectively. It is of fundamental importance for a wide range of areas such as robotics, medicine, indoor and outdoor navigation, autonomous vehicles, etc. Localization problems might be solved either with hardware or software designs. However, considering the challenging environments that would need a localization process, hardware designs might be complicated to apply to this problem. Indoor navigation can be shown as an example of these challenging environments since classical positioning methods cannot be used in such places. In this case, visual localization might be a solution since it requires either monocular or stereo images taken through the path. It would be a time-saving and cost-effective way to determine the locations that images were taken, thus the path that a robot or medical instrument, etc. take along the way. In this case, it is important the determine the performances of different localization algorithms. In this study, the performance of two different localization algorithms in indoor navigation was tested. In this context, Visual odometry and EKF SLAM algorithms were used to determine the camera trajectory utilizing the images that were taken in a straight corridor with a smartphone camera. To determine the accuracy of each method, the distances between each image-taking point were measured and compared with the distances obtained from the algorithm. Thus, root mean square error values were determined by each method. The precisions of each method were also given based on the fact that the distance between each image-taking point was equal. Therefore, the usage of both algorithms in indoor navigation was discussed.

1. Introduction

Localization is the process of determining the location of an object or person and is an interesting and challenging problem in many fields such as security, indoor navigation, robotics and interactive technologies (Haque et al., 2013). Advancements in localization-based technologies have increased the demand for applications like positioning, and real-time tracking of physical objects within buildings. Thus, there has been a marked increase in commercial interest, particularly in indoor localization services (Yassin et al., 2016). The reliable and effective operation of indoor navigation can be achieved through the collaboration of both hardware and software components. Physically tangible hardware components are systems that require careful installation, configuration, and maintenance to ensure optimal performance. Examples of popular technologies used for indoor localization include Wi-Fi, Radio Frequency Identification (RFID), ultrasonic sensors, Bluetooth, and Ultrawide Band (UWB) (Curran et al., 2011; Aguilar-Garcia et. al, 2015; Basri and El Khadimi 2016).

On the other hand, the software component is adaptable and incurs minimal costs since it is non-physical. In this context, several important localization methods are as follows: Visual Odometry, EKF SLAM, Mono SLAM, Structure from Motion (SfM), Visual Place Recognition (VPR) deep learning-based methods, etc. These techniques aim to address the challenges of autonomous navigation and mapping in complex environments, each with its advantages and limitations. Among these methods, Visual odometry is a technique for estimating the position and orientation of an agent (e.g., vehicle, human, or robot) using a series of images from one or more attached cameras. It is a cost-effective alternative to traditional methods such as Global Navigation Satellite Systems (GNSS), Inertial Navigation Systems (INS), wheel odometry, and sonar-based localization, offering greater accuracy with relative position errors typically ranging from 0.1% to 2% (Scaramuzza and Fraundorfer, 2011). In addition, Simultaneous Localization and Mapping (SLAM) is a technique that enables autonomous vehicles to simultaneously build a map of an unknown environment and determine their position. This process requires handling non-linear models due to the inherent errors in vehicle position and the relationship between the map and these errors (Sasiadek et. al, 2008). The study conducted by Leonard and Durrant-Whyte (1991) introduced the first SLAM algorithm, EKF-SLAM, which uses the Extended Kalman Filter method to solve the SLAM problem and applies a probabilistic approach to reduce the impact of sensor inaccuracies on the mobile robot's map accuracy (Naminski, 2013). The study conducted by Chatterjee et al. (2011), which tested the EKF SLAM algorithm, demonstrated that the system successfully performed localization and mapping for mobile robots in indoor environments, achieving high accuracy and producing reliable results under real-world conditions.

Localization problems might be solved with either hardware or software design in indoor environments. However, solving such a task in an indoor environment might be more complicated utilizing a different hardware design compared with the software design. In recent years visual localization algorithms have emerged as a fast and easy solution to localization problems in challenging environments. The camera images are used in visual localization to obtain orientation and position using methods such as Structure from Motion, feature matching, and deep learning to determine the path and map locations in different environments. In this context, using just visual elements such as

https://doi.org/10.5194/isprs-archives-XLVIII-2-W7-2024-183-2024 |

photographs taken with a basic camera would be sufficient to determine the path that the robot follows through the environment. In this study, the localization process was carried out using visual odometry and EKF SLAM algorithms with the images taken in a straight corridor of the building to obtain the camera movement of the camera through the environment and to determine the performances of each algorithm. The results were also compared. The real distance measurements on the ground were measured with steel tape and were the distance measurements obtained by the localization process. Thus, the accuracy of each method was presented.

2. Methods

The location information is of fundamental importance for the present communication systems that enable location-based services. The outdoor environments is relatively easy to determine the positions and locations of the objects with high accuracy utilizing the standalone cellular systems or GNSS. On the other hand, indoor applications are much more challenging in terms of localization problems since the signals of GNSS systems or cellular systems cannot work in indoor environments properly (Yassin et al., 2016).

Thus, indoor localization needs different solutions to determine the position and location of the path taken during the process. Since GNSS or cellular systems do not work properly in such environments, image-based localization techniques have come to the fore for sufficient determination of the path during the process. In this context, visual odometry and EKF-SLAM can be pointed out as the two most used methods in image-based localization studies.

2.1 Visual Odometry

Visual Odometry can be defined as the method of determining the position and motion of a camera by utilizing and analyzing a series of images. It is of fundamental importance in many different areas such as autonomous robot navigation and computer vision (Scaramuzza and Fraundorfer, 2011). Visual odometry is suitable for such applications since it utilizes consumer-grade cameras that enable a direct determination of the position of the vehicles and robots, unlike expensive sensors and systems (Gonzalez et al., 2012). Visual odometry does not provide a map of the environment, unlike SLAM which enables navigation and localization without storing observed landmarks. This method is known for its cost-effectiveness, ease of use and reliability. Visual odometry operates in environments where no external signals or references exist, making it useful in environments with weak or no GNSS signal (Galati et al., 2017). Visual odometry incrementally estimates a vehicle's motion by analysing sequential camera images and computing the relative pose between viewpoints using 2D bearing vectors derived from the captured features. At time *k*, the visual odometry algorithm takes two consecutive images, I_k and I_{k-1} s input and provides an incremental estimate of the motion relative to the local camera reference frame. This motion estimate is represented as $\delta *_{k,k-1} \in$ R^3:

$$\delta *_{k,k-1} = (\Delta s *_k, \Delta\theta_k) \quad (1)$$

$\Delta s*_k$, represents the translational movement of the camera in the 2D plane; $\Delta\theta_k$, refers to the change in orientation or rotation of the camera between two consecutive frames. One of the key challenges in visual odometry is scale ambiguity during motion, which requires the estimation of a scale factor to recover the true distance. This uncertainty can be mitigated by incorporating additional measurements, such as prior knowledge of the camera's height. Motion estimates are generated using randomly sampled correspondences, and their mean and covariance are computed. In this way, while visual odometry accurately predicts motion, it effectively manages uncertainties and error accumulation throughout the process (Ouerghi et. al., 2018).

The workflow of the visual odometry can be seen in Figure 1.

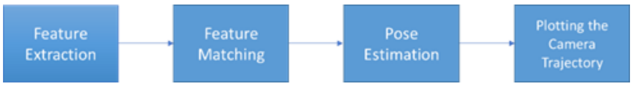

Figure 1. The workflow of visual odometry.

2.2 EKF SLAM

EKF SLAM is a widely used method for simultaneous localization and mapping (SLAM) in mobile robots. In this approach, the robot’s position and surrounding landmarks are tracked to build a map of the environment. The two fundamental equations that describe this process are the EKF state model (2) and the observation model (3) can be represented as follows:

$$X_{k+1} = f(X_k, U_k, w_k) \quad (2)$$

$$Z_{k+1} = h(X_{k+1}, v_{k+1}) \quad (3)$$

The process and observation noise are represented by w_k ~ N(0, Q_k) and v_k ~ N(0, R_k), respectively. X_{k+1} is the estimated state vector at time *k+1*, with discrete time and known input U_k. Z_{k+1} is the estimated measurement vector at *k+1*, with v_k as the observation noise. Q_k and R_k represent the covariance matrices for prediction and observation. EKF provides an approximation of the optimal state estimate, with the aim of EKF-SLAM being to recursively estimate the landmark state X_k as specified by the Z_{k+1} measurement. In addition, in EKF-SLAM, the Jacobian matrix plays a crucial role in both the prediction and update steps by linearizing nonlinear systems, thereby facilitating the prediction and observation processes. (Ullah et al., 2020).

3. Case Study

3.1 Visual Odometry Application

Visual odometry is a method that is used to obtain movement of the camera through an environment by examining consecutive image frames. In this study, the visual odometry approach is applied to estimate the pose of a camera based on consecutive images utilizing the implementation in Python with OpenCV.

We utilized consecutive images captured by a smartphone camera (Vivo Y21 S) in a straight corridor to obtain camera poses in an indoor environment. The process initially performs The ORB (Oriented FAST and Rotated BRIEF) feature detector and descriptor to extract features in these sequential images. The feature extraction step is important for subsequent image comparisons. Following feature extraction, we utilized a feature-matching strategy implementing a brute-force matcher that is fundamental for identifying correspondences between successive image frames. The Hamming distance was utilized as a metric during the matching process, enabling the accurate tracking of motion between images since it is suitable for binary descriptors. The quality of these matches is very important as they directly affect the subsequent pose estimation process. The key points and matched features between the two images can be seen in Figure 2 and Figure 3 respectively.

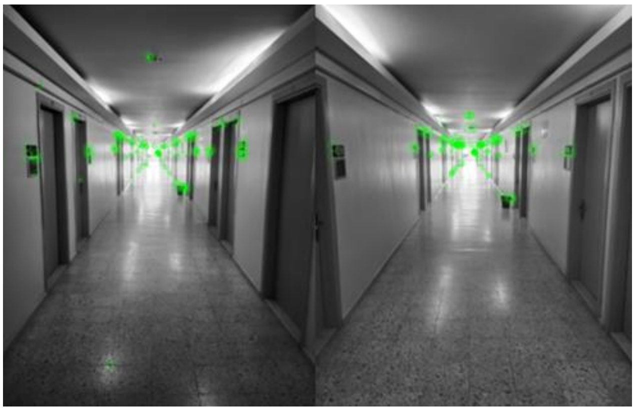

Figure 2. The key points identified in different images.

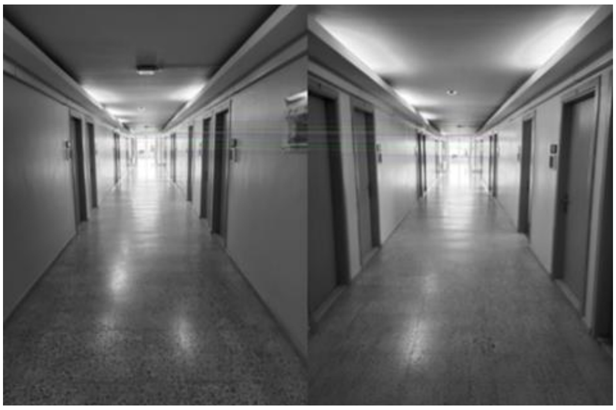

Figure 3. The matched features between two images.

After matches are identified, the algorithm determines the camera’s pose—its orientation and position—by computing the essential matrix from the matched features and then using this matrix to recover the camera’s rotation and translation. The essential matrix provides the relationship between consecutive views and therefore it is of fundamental importance to understand the path through the environment. The rotation matrix and translation vector are iteratively adjusted to improve the estimate of the camera's trajectory, starting with the identity matrix and zero translation for the initial image. The iterative process that is used in the study enables us to update the camera's pose constantly, keeping track of its trajectory through the environment. The recent image data is used to recalculate the pose in each iteration, which gives us a dynamic representation of the camera's movement over time. The camera pose information was later stored in separate lists for the x, y, and z coordinates. The trajectory is built up across a series of images, with the camera’s position being adjusted according to the estimated relative motion between each frame. Then, the result is scaled based on the known distance between each image acquisition point. The images were taken in equal intervals during the image acquisition process. A scale factor was assigned to get the real-world distances using the following formula:

$$\text{Scale Factor} = (\text{Real Distance})/(\text{Odometry Distance}) \qquad (4)$$

$$\text{P(scaled)}=\text{P(original)}*\text{Scale Factor} \qquad (5)$$

The visualization of the localization result in a 3D plot was obtained using Matplotlib. The 3D plot of the localization process can be seen in Figure 4.

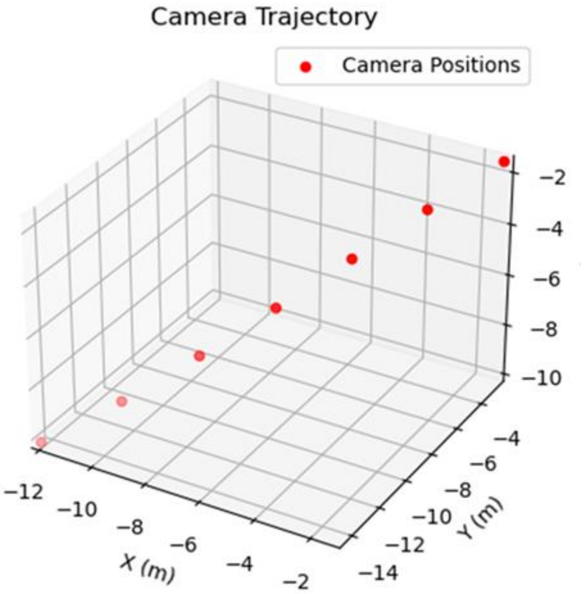

Figure 4. Localization with visual odometry.

As the last step, the accuracy assessment process was carried out based on the real-world measurement between the image acquisition points, which is 305 cm between each point. The residuals (V) were calculated by taking the differences between model measurements and the ground truth value. Thus, the root mean square error (RMSE) was calculated using Equation 6, where n is the number of measurements. The RMSE was found to be ±3.03cm with visual odometry.

$$\text{RMSE}=\pm\sqrt{(([VV]/n))} \qquad (6)$$

Where;

VV = the square of the residuals

n = number of distances

The measured distances between each image pair are shown in Table 1. As can be seen from the measurements, each distance is relatively close to the other which shows a significant precision.

Image Pair	Measured Distance (cm)
1 to 2	308.08
2 to 3	307.92
3 to 4	308.15
4 to 5	308.03
5 to 6	307.88
6 to 7	308.10

Table 1. The measured distances between each image pair with visual odometry.

The Figure 5 depicts the distance differences between the ground truth and algorithm in accordance with the RMSE value.

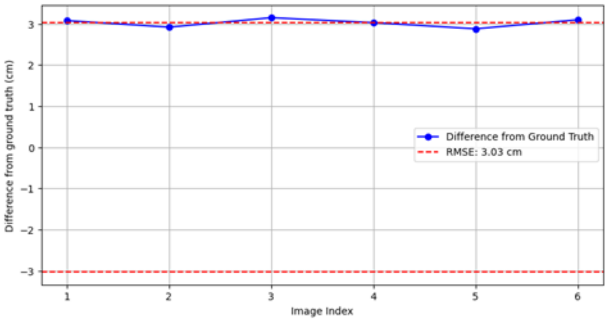

Figure 5. RMSE Value and the distance differences from ground truth between each image obtained with Visual Odometry.

The standard deviation value was also calculated using the Equation 7:

$$\sigma = \sqrt{\left(\left(\sum(X_i - \mu)^2 / n\right)\right)} \quad (7)$$

Where;

σ = Standard Deviation

Xi = Distances

μ = Mean

n = Total number of distances

The standard deviation was found to be 0.09 cm with visual odometry.

3.2 EKF SLAM Application

The Extended Kalman Filter (EKF) for Simultaneous Localization and Mapping (SLAM) used in this study utilized the ORB (Oriented FAST and Rotated BRIEF) feature detector and descriptor to extract features in the sequential images. The images were matched for each sequential image pair to obtain the relative camera pose which consists of the estimation of both the translation and rotation between the images. The key of the EKF SLAM algorithm is the estimation of the camera trajectory and modification of its orientation and position over time. The flattened representation of the rotation matrices and the 3D position of the camera were obtained by the state vector. Thus, the motion of the camera in three-dimensional space could be modelled. The covariance matrix captures the uncertainty associated with the state estimates. The predictions and updates are performed with the EKF by embedding both the noise covariance and the estimated pose matrices. These are responsible for the errors in the motion model and measurement process. This allows the filter to refine the trajectory estimates iteratively. The estimated camera positions that include X, Y, and Z coordinates are stored and visualized in a 3D plot afterward. Thus, the camera's motion through the environment is graphically represented providing an understanding of how various camera perspectives relate spatially to one another and the overall path of movement.

The result is also scaled based on the known distance that was measured between each image-taking point using the following formulas:

$$\text{Scale Factor} = (\text{Real Distance})/(\text{EKF SLAM Distance}) \quad (8)$$

$$P(\text{scaled}) = P(\text{original}) * \text{Scale Factor} \quad (9)$$

The code manages different challenges in SLAM including feature extraction and matching, pose estimation, and state estimation presenting an approach to obtain the camera movement in an indoor environment based on visual input. It is a basic but practical example of how EKF can be applied to visual SLAM problems, highlighting the key concepts and computational steps involved in creating a 3D trajectory map from image sequences. 7 images were used during the localization with the EKF SLAM algorithm. The result of the localization with the EKF SLAM algorithm can be seen in Figure 6.

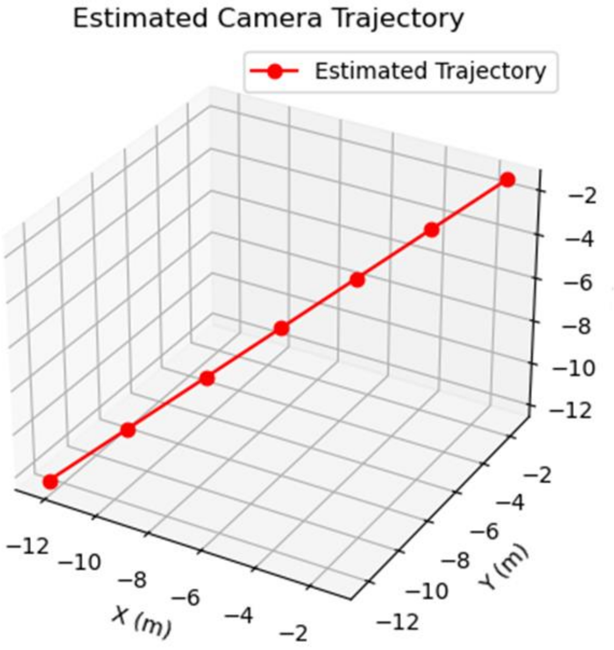

Figure 6. Localization with EKF SLAM.

https://doi.org/10.5194/isprs-archives-XLVIII-2-W7-2024-183-2024 | © Author(s) 2024. CC BY 4.0 License.

As the last step, the accuracy of the EKF SLAM was determined by comparing the ground truth distances with the distances from the algorithm.

The measured distances between each image pair can be seen in Table 2.

Image Pair	Measured Distance (cm)
1 to 2	313.2
2 to 3	313.6
3 to 4	310.5
4 to 5	315.1
5 to 6	328.2
6 to 7	328.2

Table 2. The measured distances between each image pair with EKF SLAM

RMSE value was found to be ± 15 cm. The distances between each image and the RMSE value can be seen in Figure 7.

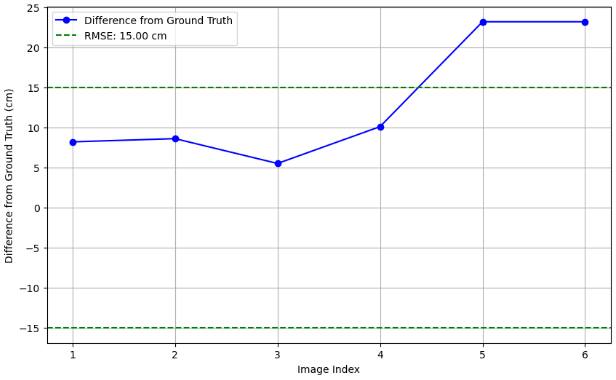

Figure 7. RMSE Value and the distance differences from ground truth between each image obtained with EKF SLAM.

The standard deviation was also calculated and it was found to be 7.24 cm with EKF SLAM.

4. Conclusion

Image-based localization methods are important for indoor navigation studies since conventional positioning methods such as GNSS are limited due to signal blockages in challenging environments. The indoor navigation problem might be solved either with hardware or algorithm designs. However, hardware designs might be more time-consuming and costlier compared to algorithm designs. In these scenarios, image-based methods have come to the fore as a solution to the navigation problem in indoor environments. The usage of images enables the acquisition of visual information, therefore the determination of the position and orientation of the path taken is possible by utilizing such methods. In this case, image-based localization can be useful in shopping malls, hospitals, airports, etc., where positioning is important for safety, accessibility, and efficiency. The advancements in visual localization algorithms such as SLAM and Visual Odometry make it possible to obtain the position and orientation of the path taken during the process. This technology improves the efficiency of facility management and user experiences.

In this study, the performance of two visual localization algorithms in indoor navigation was evaluated. In this context, 7 monocular photographs in total were captured in a straight corridor of a building with a smartphone camera. The localization process was carried out in Python environment using visual odometry and EKF SLAM methods. The accuracy of each method was determined by comparing the real-world measurements between each point and the distances obtained from the algorithm. The RMSE of visual odometry was found to be ±3.03 cm, while the accuracy of the EKF SLAM was found to be ±15 cm. The standard deviation values were also calculated and were found to be 0.09 cm for visual odometry while it was found 7.24 cm for EKF SLAM. It has been seen that visual odometry gives more precise results in comparison to EKF SLAM. It has been seen that in the case of using monocular images, visual odometry might give more satisfactory results but since each algorithm gave the local position of the movement, both can be used for basic applications. The use of different sensors and stereo-view images might contribute to the accuracy as well. In future studies, it is planned to investigate the monocular vs. stereo view in indoor localization studies. On the other hand, deep learning-based methods can also be applied to indoor navigation problems in future studies.

References

Aguilar-Garcia, A., Fortes, S., Colin, E., & Barco, R., 2015. Enhancing RFID indoor localization with cellular technologies. EURASIP *Journal on Wireless Communications and Networking*, 2015, 1-22.

Basri, C., El Khadimi, A., 2016. Survey on indoor localization system and recent advances of WIFI fingerprinting technique. IEEE 5th International conference on multimedia computing and systems (ICMCS), 253-259.

Chatterjee, A., Ray, O., Chatterjee, A., & Rakshit, A., 2011. Development of a real-life EKF based SLAM system for mobile robots employing vision sensing. Expert Systems with Applications, 38(7), 8266-8274.

Curran, K., Furey, E., Lunney, T., Santos, J., Woods, D., & McCaughey, A., 2011. An evaluation of indoor location determination technologies. *Journal of Location Based Services*, 5(2), 61-78.

Galati, R., Reina, G., Messina, A., Gentile, A., 2017. Survey and navigation in agricultural environments using robotic technologies. 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, pp. 1–6.

Gonzalez, R., Rodriguez, F., Guzman, J. L., Pradalier, C., Siegwart, R., 2012. Combined visual odometry and visual compass for off-road mobile robots localization. *Robotica* 30 (6), 865–878.

Haque, I. T., Assi, C., 2013. Profiling-based indoor localization schemes. *IEEE Systems Journal*, 9(1), 76-85.

Leonard, J. J., Durrant-Whyte, H. F., 1991. Mobile robot localization by tracking geometric beacons. IEEE Transactions on Robotics and Automation, 7(3), 376-382.

Naminski, Megan R., 2013. "An Analysis of Simultaneous Localization and Mapping (SLAM) Algorithms". Mathematics, Statistics, and Computer Science Honors Projects. 29.

Ouerghi, S., Boutteau, R., Savatier, X., & Tlili, F., 2018. Visual odometry and place recognition fusion for vehicle position tracking in urban environments. Sensors, 18(4), 939.

Sasiadek, J. Z., Monjazeb, A., & Necsulescu, D., 2008. Navigation of an autonomous mobile robot using EKF-SLAM and FastSLAM. IEEE 16th Mediterranean Conference on Control and Automation, 517-522).

Scaramuzza, D., Fraundorfer, F., 2011. Tutorial: Visual Odometry. IEEE Robotics & Automation Magaz., 18(4), 80-92.

Ullah, I., Su, X., Zhang, X., & Choi, D., 2020. Simultaneous localization and mapping based on Kalman filter and extended Kalman filter. Wireless Communications and Mobile Computing, 2020(1), 2138643.

Yassin, A., Nasser, Y., Awad, M., Al-Dubai, A., Liu, R., Yuen, C., Raulefs, R. & Aboutanios, E., 2016. Recent advances in indoor localization: A survey on theoretical approaches and applications. IEEE Communications Surveys & Tutorials, 19(2), 1327-1346.

https://doi.org/10.5194/isprs-archives-XLVIII-2-W7-2024-183-2024