A Tightly-coupled Semantic SLAM System with Visual, Inertial and Surround-view Sensors for Autonomous Indoor Parking

Xuan Shao
School of Software Engineering, Tongji Unversity
Shanghai, China
1810553@tongji.edu.cn

Lin Zhang*
School of Software Engineering, Tongji Unversity
Shanghai, China
cslinzhang@tongji.edu.cn

Tianjun Zhang
School of Software Engineering, Tongji Unversity
Shanghai, China
1911036@tongji.edu.cn

Ying Shen*
School of Software Engineering, Tongji Unversity
Shanghai, China
yingshen@tongji.edu.cn

Hongyu Li
Tongdun AI Institute
Shanghai, China
hongyu.li@tongdun.net

Yicong Zhou
Department of Computer and Information Science,
University of Macau, China
yicongzhou@um.edu.mo

ABSTRACT

The semantic SLAM (simultaneous localization and mapping) system is an indispensable module for autonomous indoor parking. Monocular and binocular visual cameras constitute the basic configuration to build such a system. Features used in existing SLAM systems are often dynamically movable, blurred and repetitively textured. By contrast, semantic features on the ground are more stable and consistent in the indoor parking environment. Due to their inabilities to perceive salient features on the ground, existing SLAM systems are prone to tracking loss during navigation. Therefore, a surround-view camera system capturing images from a top-down viewpoint is necessarily called for. To this end, this paper proposes a novel tightly-coupled semantic SLAM system by integrating *V*isual, *I*nertial, and *S*urround-view sensors, VIS_{SLAM} for short, for autonomous indoor parking. In VIS_{SLAM}, apart from low-level visual features and IMU (inertial measurement unit) motion data, parking-slots in surround-view images are also detected and geometrically associated, forming semantic constraints. Specifically, each parking-slot can impose a surround-view constraint that can be split into an adjacency term and a registration term. The former pre-defines the position of each individual parking-slot subject to whether it has an adjacent neighbor. The latter further constrains by registering between each observed parking-slot and its position in the world coordinate system. To validate the effectiveness and efficiency of VIS_{SLAM}, a large-scale dataset composed of synchronous multi-sensor data collected from typical indoor parking sites is established, which is the first of its kind. The collected dataset has been made publicly available at https://cslinzhang.github.io/VISSLAM/.

*Corresponding Authors: Lin Zhang and Ying Shen

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413867

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**.

KEYWORDS

autonomous indoor parking, surround-view camera system, parking-slot, semantic mapping

ACM Reference Format:
Xuan Shao, Lin Zhang, Tianjun Zhang, Ying Shen, Hongyu Li, and Yicong Zhou. 2020. A Tightly-coupled Semantic SLAM System with Visual, Inertial and Surround-view Sensors for Autonomous Indoor Parking. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413867

1 INTRODUCTION

In the autonomous driving industry, self-parking is a problem that needs to be solved urgently to achieve the "last-mile" navigation [22]. It enables the autonomous vehicle to successfully maneuver in an unknown indoor parking environment. When an autonomous vehicle navigates in such an indoor environment, an important thing is to sense and understand its surroundings. In order to achieve this goal, a SLAM system [5] that simultaneously estimates the vehicle motion as well as constructs the map of its surrounding environment is necessarily called for.

Monocular and binocular visual cameras constitute the basic sensor configuration to build such a SLAM system. Often, visual cameras and IMU can collaborate further to establish a VI-SLAM (visual-inertial SLAM) system to prevent tracking loss during navigation. The visual camera operates stably in texture-rich scenes, whereas the IMU estimates motion of the vehicle by directly measuring its angular velocity and linear acceleration, thereby complementing the camera in an environment of severe jitters or missing textures. However, VI-SLAM systems usually depend on low-level visual features, such as points and lines, or directly utilize the intensity of pixels. Both low-level visual features and image pixels suffer from instability and inconsistency during tracking in indoor parking environment. For example, when the vehicle is making a

turn, it may face a low-textured wall, leading to tracking loss. Besides, these SLAM systems fail to understand semantic information around the vehicle.

In order to help self-driving cars understand their surroundings, a semantic SLAM system needs to incorporate semantic information (e.g., cars or people) to localize itself and to construct the map [2–4, 21, 25, 26]. Unfortunately, for indoor parking, commonly studied semantic objects, such as cars and people, are generally dynamic, which serve little to or even compromise the localization and mapping accuracy. By contrast, for this specific application scenario, the parking-slot painted on the ground is a kind of stable, salient, and easy-to-detect semantic feature. At present, unfortunately, few eminent semantic SLAM systems have made use of parking-slot information effectively.

Taking aforementioned analysis into considerations, in this paper, we attempt to build a semantic SLAM system, namely VIS_{SLAM} (a SLAM system integrating visual, inertial, and surround-view sensors), specially for the task of autonomous indoor parking. Our contributions can be summarized as follows:

(1) Specially designed for navigation in the indoor parking site, VIS_{SLAM} is the first tightly-coupled semantic SLAM system that fully explores parking-slots detected in surround-views in its optimization framework. Surround-view images are synthesized online from bird's-eye views generated from a surround-view camera system, comprising four fisheye cameras. Since VIS_{SLAM} can construct maps with semantic parking-slot information, it can be naturally integrated into a high-level self-parking system.
(2) In order to improve the localization accuracy and to construct the semantic map, in VIS_{SLAM} parking-slots in surround-view images are leveraged for optimization in which they are modeled as an adjacency term subject to the existence of adjacent neighbors and a registration term constraining by registering between the observed parking-slots and their positions in the world coordinate system. Experiments demonstrate that the semantic constraints induced by parking-slots can significantly improve the performance of VIS_{SLAM}.
(3) At present, there is no publicly available dataset for SLAM research that contains visual information of surround-views. To fill this gap to some extent, in this work we established a large-scale dataset comprising synchronous multi-sensor data from typical indoor parking sites. It will benefit further SLAM studies, especially conducted for autonomous indoor parking.

The remainder of this paper is organized as follows. Sect. 2 introduces the related work. Sect. 3 presents the overall framework of VIS_{SLAM}. Details for sensor calibration and system's implementation are presented in Sect. 4 and Sect. 5, respectively. Sect. 6 reports the experimental results and Sect. 7 concludes the paper.

2 RELATED WORK

2.1 VI-SLAM Systems

VI-SLAM systems can be roughly categorized as loosely-coupled and tightly-coupled ones according to their ways of sensor fusion. Typical studies belonging to the former ones include [12] and [23]. In [12], Munguía *et al.* fused measurements from different sensors and the system's optimization was performed based on Kalman filtering in a loosely-coupled manner. In [23], Weiss and Siegwart estimated the full and metric scaled state of a camera-IMU device in real time by decoupling of the visual pose estimate and the filter state estimation. Since both the systems in [12, 23] separately estimate the motion of the IMU and the camera, they fail to obtain highly consistent localization results due to the lack of complementary information.

By contrast, the latter ones are currently popular schemes adopted in the field of autonomous driving, which combine the state of the IMU and the camera to perform state estimation to ensure a consistent localization result [9–11, 15, 17, 18, 20, 29]. Eminent studies along this technical line are briefly reviewed here. MSCKF (multi-state constraint Kalman filter) is a real-time visual-inertial navigation system based on EKF (extended Kalman filter) [11]. It can provide accurate poses in large-scale environments and the time complexity of the MSCKF algorithm is only related to the number of features. However, the back-end of MSCKF is based on Kalman filter in which global information cannot be explored for optimization. OKVIS (open keyframe-based visual-inertial SLAM) [9] predicts the current state based on IMU measurement, and performs feature extraction and feature matching based on prediction. But it does not support relocation, and there is no loop-closing. In [15], Mur-Artal and Tardos incorporated IMU measurement into the ORB-SLAM system [7, 13, 14, 16], which is widely used in the field of autonomous driving. VINS (a monocular visual-inertial system) [17, 18] is a robust and versatile monocular visual-inertial state estimator. Its front-end resorts to KLT (Kanade-Lucas-Tomasi) tracker [1] to track Harris corner points [8], and its back-end makes use of sliding windows for optimization. The pre-integration of the IMU and the vision-IMU alignment ensure VINS' robustness and stability. S-MSCKF (stereo multi-state constraint Kalman filter) [20] is a binocular version of MSCKF, resorting to Fast corner [19] and KLT tracker [1] for tracking. PIRVS (PerceptIn robotics vision system) [29] tightly couples vision sensors and IMUs while loosely coupling with other sensors. It needs to be noted that maps constructed by these VI-SLAM systems only provide geometric information, lacking of a semantic understanding of the environment.

2.2 Semantic SLAM Systems

In order to acquire a semantic understanding of the environment, recent studies have begun to incorporate semantic features to SLAM systems [4, 21, 25, 30]. VINet (a sequence-to-sequence learning approach to visual-inertial odometry) [4] is an end-to-end VIO (visual-inertial odometry) that integrates deep learning and sensor fusion. But it does not have loop-closing and map construction components, so it is actually not a complete VI-SLAM system. CNN-SLAM [21] exploits a CNN (convolutional neural network) to estimate the depth of a single image and resorts to a semi-dense direct method to produce the final globally consistent map. In [25], Yang *et al.* extracted planar features from a 3D plane model and applied them to SLAM systems in a low-texture environment. Note that features used in these semantic SLAM systems are usually dynamically movable, blurred and repetitively textured. By contrast, parking-slots on the ground embody the stable and consistent information in

the indoor parking environment. Due to their inabilities to perceive salient features on the ground, the aforementioned SLAM systems are prone to tracking loss during navigation. To the best of our knowledge, the latest work that leverages features detected on the ground is the one established in [30]. In [30], Zhao *et al.* detected parking-slots in the surround-view images and incorporated them to the SLAM system they built. However, artificial landmarks were used to facilitate localization in Zhao *et al.*'s system, whereas parking-slots contributed little for optimization.

3 VIS_{SLAM}

The overall framework of VIS_{SLAM} is shown in Fig. 1. Sensor configuration of VIS_{SLAM} consists of a front-view camera, an IMU and four fisheye cameras facing ground to form a surround-view camera system. Visual features from the front-view camera, pre-integrated IMU measurements between two consecutive keyframes and parking-slots from the surround-view camera system constitute the multi-modal sensor data for VIS_{SLAM}. There are two major components in VIS_{SLAM}, sensor calibration and joint optimization. Sensor calibration is responsible for multi-modal sensor data fusion, which will be introduced in Sect. 4. The joint optimization model plays a critical role in tightly fusing multi-modal sensor measurements, which is the core of VIS_{SLAM}. Its details will be thoroughly presented in this section with regard to its formulation and all error terms during optimization.

3.1 Joint Optimization Model Formulation

Given keypoints $\mathcal{Z}$ in the front-view image, parking-slot observations O in the surround-view image and IMU measurements $\mathcal{M}$, the proposed joint optimization model for VIS_{SLAM} determines optimal camera poses $\mathcal{T}$, map points $\mathcal{P}$ matched with $\mathcal{Z}$ as well as parking-slot locations $\mathcal{L}$, jointly. Such an optimization problem can be defined as,

$$\{\mathcal{L},\mathcal{T},\mathcal{P}\}^* = \arg\max_{\mathcal{L},\mathcal{T},\mathcal{P}} p(\mathcal{L},\mathcal{T},\mathcal{P}|O,\mathcal{Z},\mathcal{M}). \tag{1}$$

Further, We reformulate p with Bayes' theorem as,

$$\begin{aligned} p(\mathcal{L},\mathcal{T},\mathcal{P}|O,\mathcal{Z},\mathcal{M}) &= \frac{p(\mathcal{L},\mathcal{T},\mathcal{P})p(O,\mathcal{Z},\mathcal{M}|\mathcal{L},\mathcal{T},\mathcal{P})}{p(O,\mathcal{Z},\mathcal{M})} \\ &\propto p(\mathcal{L},\mathcal{T},\mathcal{P})p(O,\mathcal{Z},\mathcal{M}|\mathcal{L},\mathcal{T},\mathcal{P}). \end{aligned} \tag{2}$$

Since keypoints $\mathcal{Z}$ and parking-slot observations O are independently observed by two sensor modalities, p can be factorized by separating parking-slot observations from other measurements, i.e.,

$$\begin{aligned} &p(\mathcal{L},\mathcal{T},\mathcal{P}|O,\mathcal{Z},\mathcal{M}) \\ &\propto p(\mathcal{L})p(\mathcal{T},\mathcal{P})p(O|\mathcal{L},\mathcal{T},\mathcal{P})p(\mathcal{Z},\mathcal{M}|\mathcal{L},\mathcal{T},\mathcal{P}) \\ &= p(\mathcal{L})p(\mathcal{T},\mathcal{P})p(O|\mathcal{L},\mathcal{T})p(\mathcal{Z},\mathcal{M}|\mathcal{T},\mathcal{P}) \\ &= \underbrace{p(\mathcal{T},\mathcal{P})p(\mathcal{Z},\mathcal{M}|\mathcal{T},\mathcal{P})}_{visual-inertial\ term} \underbrace{\overbrace{p(\mathcal{L})}^{prior}\overbrace{p(O|\mathcal{L},\mathcal{T})}^{observation}}_{surround-view\ term}, \end{aligned} \tag{3}$$

where the first two terms are with visual features and IMU motion data, and the latter is the surround-view error term. Concretely, following [15], the visual-inertial term can be converted into a visual error term and an inertial error term, $\mathbf{E}_V$ and $\mathbf{E}_I$, respectively. $\mathbf{E}_V$ links each keypoint and its projecting map point while $\mathbf{E}_I$ constrains consecutive keyframes by visual-inertial alignment, predicting stable and reliable camera poses estimation and map point locations. Parking-slots in surround-view images encode abundant information, the location, the width, the detection confidence, and the adjacency property *et al.*, imposing a surround-view constraint $\mathbf{E}_S$. Therefore, in order to find out optimal estimation, we jointly optimize visual, inertial and surround-view error terms in a tightly-coupled objective,

$$\{\mathcal{L},\mathcal{T},\mathcal{P}\}^* = \arg\min_{\mathcal{L},\mathcal{T},\mathcal{P}} \mathbf{E}_V + \mathbf{E}_I + \mathbf{E}_S. \tag{4}$$

Intuitively, with Eq. 4, VIS_{SLAM} is optimized by jointly minimizing errors of visual re-projection error, IMU motion error and surround-view error over parking-slots. The model of Eq. 4 is in charge of dealing with both low-level geometric/motion data as well as semantic features in the surround-view image, simultaneously. It enables robust perception of indoor parking environment, avoiding vulnerability to blur, dramatic lighting changes, and low-texture conditions as in the traditional SLAM system. Three error terms of Eq. 4, $\mathbf{E}_V$, $\mathbf{E}_I$, and $\mathbf{E}_S$, are detailed in the subsequent subsections.

3.2 Visual Error Term

The visual error term ${}_v\mathbf{e}_{kn}$ involving the n-th map point $\mathbf{P}_n$ and the front-view camera pose $\mathbf{T}_k \in SE(3)$ of the k-th keyframe is defined as the reprojection error with respect to the matched observation $\mathbf{z}_k^n$, i.e.,

$${}_v\mathbf{e}_{kn} = \mathbf{z}_k^n - \phi_k(\mathbf{T}_k, \mathbf{P}_n), \tag{5}$$

where $\phi_k(\cdot)$ is the projection function of the front-view camera at the time when taking the k-th keyframe. Given the set of camera poses $\mathcal{T} = \{\mathbf{T}_k\}_{k=1}^K$ and map points $\mathcal{P} = \{\mathbf{P}_n\}_{n=1}^N$, $\mathbf{E}_V$ tackles the problem of jointly optimizing camera poses $\mathcal{T}$ and map points $\mathcal{P}$, i.e.,

$$\mathbf{E}_V = \sum_{k=1}^{K}\sum_{n=1}^{N} \rho_h({}_v\mathbf{e}_{kn}^T \, \Lambda_{kn}^{-1} \, {}_v\mathbf{e}_{kn}), \tag{6}$$

where $\rho_h(\cdot)$ is the Huber kernel function for robustness to outliers and $\Lambda_{kn} = \sigma_{kn}^2\mathbf{I}_{2\times 2}$ is covariance matrix associated to the scale at which the keypoint is detected.

3.3 IMU Error Term

The motion (orientation, velocity, position) between two consecutive keyframes can be determined by either pre-integrated IMU data or the visual odometry. Each IMU error term ${}_m\mathbf{e}_{ij}$ links the i-th and the j-th keyframes, i.e.,

$${}_m\mathbf{e}_{ij} = [{}_R\mathbf{e}_{ij} \; {}_V\mathbf{e}_{ij} \; {}_P\mathbf{e}_{ij}], \tag{7}$$

where ${}_R\mathbf{e}_{ij}$, ${}_V\mathbf{e}_{ij}$, ${}_P\mathbf{e}_{ij}$ denote the orientation, the velocity, and the position error terms between consecutive keyframes, respectively. Each error term is defined as the difference between IMU and visual measurements. Thus, $\mathbf{E}_I$ is defined as,

$$\mathbf{E}_I = \sum_{i=1}^{K} \rho_h({}_m\mathbf{e}_{ij}^{-1} \; \Sigma_i \; {}_m\mathbf{e}_{ij}), \tag{8}$$

where Σ_i is the information matrix according to [15].

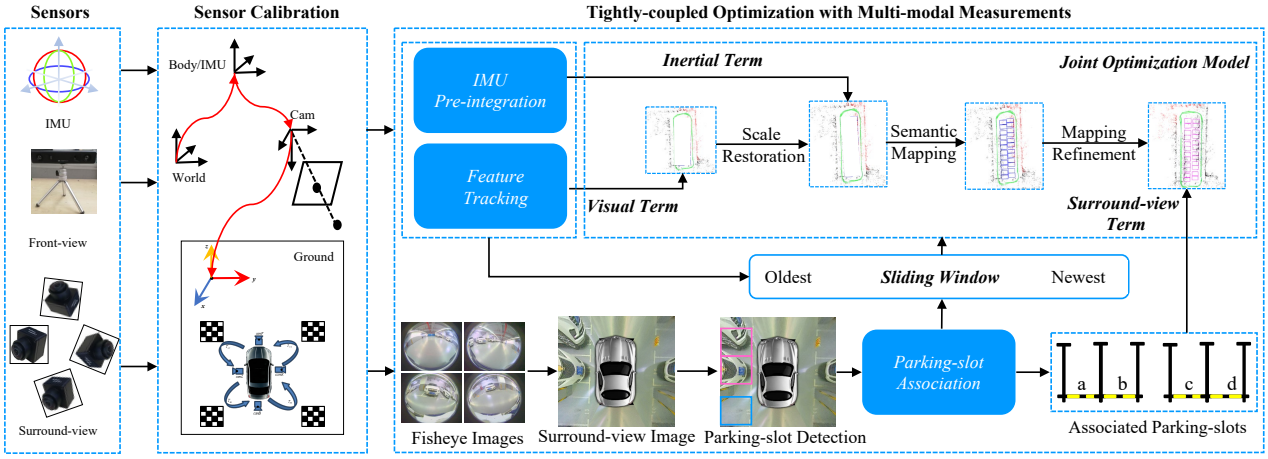

Figure 1: The overall processing pipeline of VIS_{SLAM}. Multi-modal sensors are first spatially registered with one another. Visual features are detected and tracked to construct a 3D map of an indoor parking site with no scale. By aligning pre-integrated IMU measurement with the visual features in the front-view image, a map with metric scale can be obtained. In order to build a semantic map suitable for autonomous indoor parking, parking-slots in each surround-view image are detected and geometrically associated to constitute a surround-view constraint. The visual term, IMU term as well as the surround-view term are integrated into VIS_{SLAM} during optimization. Joint optimization is performed in a sliding window, giving a trade-off between the speed and the flexibility.

3.4 Surround-view Error Term

According to Eq. 3, the surround-view error term $\mathbf{E}_S$ is split into a prior error term and an observation error term corresponding to $p(L)$ and $p(O|T,L)$ respectively. (Refer to Fig. 2). The prior error term is denoted by $\mathbf{E}_{Adj}$. It predefines the position of each individual parking-slot subject to whether it has a neighboring parking-slot. The observation error term $\mathbf{E}_{Reg}$ further constrains by registering between each observation and its position in the world coordinate system. Therefore, $\mathbf{E}_S$ can be defined as,

$$\mathbf{E}_S = \mathbf{E}_{Adj} + \mathbf{E}_{Reg}. \tag{9}$$

3.4.1 Notation. Assuming that there exist M parking-slots in the indoor parking site. Each parking-slot is represented by two marking-points ($\mathbf{p}_1$ and $\mathbf{p}_2$ in Fig. 2). We denote positions of all parking-slots by $\mathcal{L} = \{\mathbf{L}_m\}_{m=1}^{M}$, $\mathbf{L}_m \in \mathcal{R}^{3\times1}$, each of which is defined as the mid-point of the entrance line connecting the two marking-points. Additionally, widths of all parking-slots can be computed as lengths of all entrance lines and are denoted by $\mathcal{W} = \{\mathbf{W}_m\}_{m=1}^{M}$, $\mathbf{W}_m \in \mathcal{R}^{3\times1}$. At time t, the vehicle obtains K_t parking-slot observations, denoted by $O_t = \{\mathbf{O}_t^1; \mathbf{O}_t^2; ...; \mathbf{O}_t^{K_t}\}$. Parking-slots associations are denoted by $\dagger_t = \{y_t^1; y_t^2; ...; y_t^{K_t}\}$, where $y_t^i \in \{1; ...; M\}$. For example, at time $t = 1$, the surround-view camera system obtains three measurements $O_1 = \{\mathbf{O}_1^1; \mathbf{O}_1^2; \mathbf{O}_1^3\}$. And these three measurements are from parking-slots No. 2, No. 3 and No. 4, then $\dagger_1 = \{y_1^1; y_1^2; y_1^3\} = \{2; 3; 4\}$.

3.4.2 Adjacency Term. $p(\mathcal{L})$ models the prior distributions for positions of all parking-slots, each of which is independent with one another, i.e.,

$$p(\mathcal{L}) = \prod_{t=1}^{T}\prod_{k=1}^{K_t} p(\mathbf{L}_{y_t^k}), \tag{10}$$

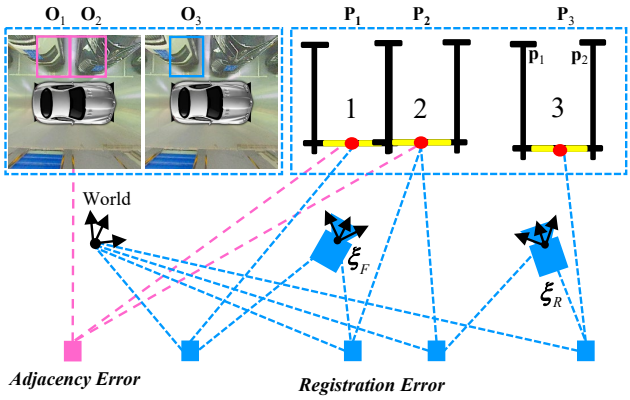

Figure 2: A surround-view error term consists of adjacency and registration error terms. The adjacency term constrains a parking-slot to closely contact its neighbor, providing the prior for the parking-slot position. The registration term finetunes the camera pose and the parking-slot position by registering between the observed parking-slot and its position in the world coordinate system.

where $p(\mathbf{L}_{y_t^k})$ is the prior of the parking-slot position associated with the k-th parking-slot observation at time t. It is defined subject to whether the parking-slot has adjacent neighbors. Take one adjacent neighbor for instance, we have the follow equation, i.e.,

$$p(\mathbf{L}_{y_t^k}) = \begin{cases} \mathcal{U} & Adj(y_t^k) \notin \dagger_t \\ \mathcal{N}(\mathbf{d}_{y_t^k} + \mathbf{L}_{Adj(y_t^k)}, \Lambda_{kt}) & Adj(y_t^k) \in \dagger_t, \end{cases} \tag{11}$$

where $\mathcal{U}$ is a uniform distribution, $\mathcal{N}$ represents a normal distribution, and $Adj(y_k^t)$ denotes the ID of the neighboring parking-slot.

$\mathbf{L}_{Adj(y_k^t)}$ represents the position of the neighboring parking-slot. Λ_{kt} models the uncertainty. $\mathbf{d}_{y_t^k}$ is a vector defined by two adjacent parking-slots as,

$$\begin{cases} \mathbf{d}_{y_t^k} & // \quad \mathbf{L}_{Adj(y_k^t)}\mathbf{L}_{y_t^k} \\ ||\mathbf{d}_{y_t^k}||_2^2 & = \quad \frac{1}{2}(\mathbf{W}_{y_t^k} + \mathbf{W}_{Adj(y_k^t)}). \end{cases} \tag{12}$$

$\mathbf{d}_{y_t^k}$ points from $\mathbf{L}_{Adj(y_k^t)}$ to $\mathbf{L}_{y_t^k}$. Intuitively, if a parking-slot sits alone with no neighbor, the distribution of its location is uniform. Otherwise, it is constrained by its neighbor to maintain the adjacency structure. Hence, the adjacency error term $\mathbf{e}_{adj}^{k,t}$ of the k-th parking-slot observed at time t is defined as,

$$\mathbf{e}_{adj}^{k,t} = \begin{cases} \mathbf{0} & Adj(y_t^k) \notin \dagger_t \\ \mathbf{d}_{y_t^k} - (\mathbf{L}_{y_t^k} - \mathbf{L}_{Adj(y_k^t)}) & Adj(y_t^k) \in \dagger_t. \end{cases} \tag{13}$$

Therefore, minimizing the adjacency error term implies iteratively tweaking each parking-slot to closely contact its adjacent neighbor.

3.4.3 Registration Term. Considering all camera poses and parking-slots, the observation term $p(O|\mathcal{T}, \mathcal{L})$ is defined as,

$$p(O|\mathcal{T}, \mathcal{L}) = \prod_{t=1}^{T}\prod_{k=1}^{K_t} p(\mathbf{O}_t^k|\mathbf{T}_t, \mathbf{L}_{y_t^k}), \tag{14}$$

where $\mathbf{T}_t$ is the camera pose at time t and $\mathbf{O}_t^k$ represents the k-th observation at time t. $p(\mathbf{O}_t^k|\mathbf{T}_t, \mathbf{L}_{y_t^k})$ is the observation probability of the k-th parking-slot observation at time t. Since each parking-slot is associated with multiple observations, it constitutes a registration problem between each observed parking-slot and its position in the world coordinate system, i.e.,

$$p(\mathbf{O}_t^k|\mathbf{T}_t, \mathbf{L}_{y_t^k}) = \mathcal{N}(\mathbf{T}_t\mathbf{L}_{y_t^k}, \Phi_{k,t}), \tag{15}$$

where $\Phi_{k,t}$ models the uncertainty. Therefore, the registration error term of the k-th parking-slot observed at time t can be defined as,

$$\mathbf{e}_{reg}^{k,t} = \mathbf{T}_t\mathbf{L}_{y_t^k} - \mathbf{O}_t^k. \tag{16}$$

3.4.4 Surround-view Error Term. Combining both the adjacency term and the registration term, the surround-view error term $\mathbf{E}_S$ can be constructed by adding up all parking-slot observations during navigation, i.e.,

$$\begin{aligned} \mathbf{E}_S &= \mathbf{E}_{Adj} + \mathbf{E}_{Reg} \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K_t} (\mathbf{e}_{adj}^{k,t})^{-1}\Lambda_{k,t}\mathbf{e}_{adj}^{k,t} + (\mathbf{e}_{reg}^{k,t})^{-1}\Phi_{k,t}\mathbf{e}_{reg}^{k,t}, \end{aligned} \tag{17}$$

where both $\Lambda_{k,t}$ and $\Phi_{k,t}$ are in proportion to the detection confidence of each parking-slot. By minimizing Eq. 17, intuitively, the objective of our proposed surround-view error term encourages both geometric and observational consistency.

4 SENSOR CALIBRATION

The configuration of VIS_{SLAM} comprises a front-view camera, an IMU and four fisheye cameras to form a surround-view camera system. For the best performance in sensor fusion, these different sensors must be spatially registered with respect to one another (Refer to Fig. 1 for details). The intrinsics of all visual sensors and the IMU can be acquired according to [24, 28]. The extrinsic calibrations can be categorized into two respects, surround-view camera system calibration and camera-IMU calibration. The former can be performed by [27]. For camera-IMU calibration, the front-view camera and IMU are considered rigidly attached and the transformation between their coordinate systems can be denoted by $\mathbf{T}_{CB}$. Specifically, we collect a set of data typically over several minutes as the camera-IMU is waved in front of a static calibration pattern. Following [6], $\mathbf{T}_{CB}$ can be then computed by optimizing the error term between IMU and camera measurement. With the camera pose $\mathbf{T}_{CW}$ obtained from the visual odometry, IMU motion in the world coordinate system $\mathbf{T}_{BW}$ can be computed as $\mathbf{T}_{BW} = \mathbf{T}_{CB}^{-1}\mathbf{T}_{CW}$. Additionally, by selecting four points $\mathcal{P}_G$ on a calibration site, the transformation $\mathbf{T}_{FG}$ from the front-view camera to the ground can be estimated by solving a PnP (Perspective-n-Point) problem between $\mathcal{P}_G$ and corresponding image pixels in each camera.

5 SYSTEM IMPLEMENTATION

5.1 Parking-slot Detection

We adopt the CNN-based approach, namely DeepPS, to detect parking-slots in a surround-view image. It first uses a CNN to detect marking-points and then uses another CNN to classify local image patterns determined by marking-point pairs. Its training details can be found in [27]. In addition, after checking whether two parking-slots are sharing the same marking-point, their adjacency property can be obtained.

5.2 Window Optimization

VIS_{SLAM} is optimized by minimizing a combination of an IMU error term, a visual error term and a surround-view error term (Refer to Fig. 3). The visual error term links map points and camera poses, whereas the IMU error term links motion data (pose, velocity and biases) between consecutive keyframes. Additionally, the surround-view error term optimizes each parking-slot and the camera pose at which the parking-slot is observed. In order to make a good trade-off between speed and flexibility, the optimization is performed within a sliding window. Frames with sufficient features and large parallax are selected as keyframes and are inserted into the sliding window. Note that there are additional states of parking-slots in the surround-view image. Therefore, frames that don't hold enough features will, nevertheless, be regarded as a new keyframe if parking-slots are detected in the corresponding surround-view image. When a new keyframe is inserted into the sliding window, it optimizes the last N keyframes in the local window and all points seen by those N keyframes. In addition, parking-slots are also incorporated during optimization. A suitable local window size has to be chosen for real-time performance. All other keyframes that share observations of map points and parking-slots contribute to the total cost but are fixed in a fixed window during optimization in order to provide a deterministic solution. The keyframe N+1 is always included in the fixed window as it constrains the IMU states. If the total number of keyframes exceeds the local window size, redundant keyframes are discarded. Since parking-slots in the surround-view image act as consistent semantic features for autonomous indoor parking, keyframes with parking-slots in the corresponding surround-view images will not be discarded.

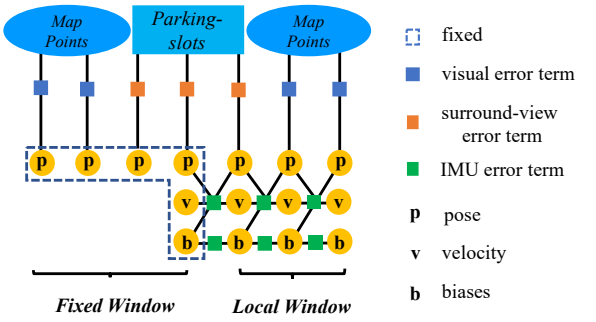

Figure 3: Sliding window optimization of VIS_{SLAM}. Local map points and parking-slots are visible in the local window. Cubes with different colors represent error terms linking corresponding variables. Frames in the local window will be optimized, whereas frames in the fixed window only contribute to the cost but are not optimized.

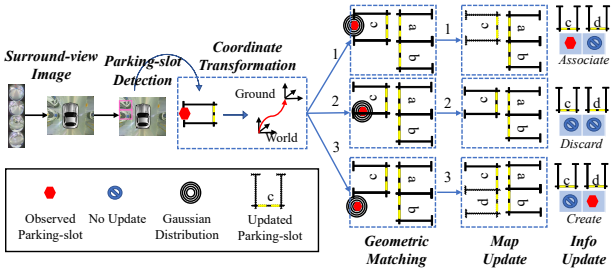

Figure 4: Parking-slot association. Parking-slot association is based on geometric distances between parking-slots in the map and the observed one. According to the distances, the parking-slot observation will be (1) associated with one in the map, (2) discarded as abnormal observation, or (3) regarded as a new one.

5.3 Parking-slot Association

The purpose of parking-slot association is to associate parking-slot observations along navigation. Since appearances on the ground within each parking-slot region are either blurred or occluded by the movable car, it is hard to distinguish parking-slots by comparing their appearances. Therefore, the parking-slot association is mainly based on geometric matching (Refer to Fig. 4 for details).

5.3.1 Geometric Association for Parking-slot. In particular, the probability distribution p_t^i of the i-th parking-slot's position detected at time t follows a Gaussian distribution, i.e.,

$$p_t^i = \mathcal{N}({}_w\mathbf{P}_{y_t^k}, \sigma), \tag{18}$$

where ${}_w\mathbf{P}_{y_t^k} = \mathbf{T}_{CW}^{-1}\mathbf{O}_t^k$ is the estimated parking-slot position at time t. $\mathbf{T}_{CW}$ is the camera pose returned by the visual odometry. σ is the information matrix. The probability of the observation associated with the j-th parking-slot in the map can be defined as,

$$f_t^i(j) = p_t^i(\mathbf{L}_j), \tag{19}$$

where $\mathbf{L}_j$ denotes the position of the j-th parking-slot.

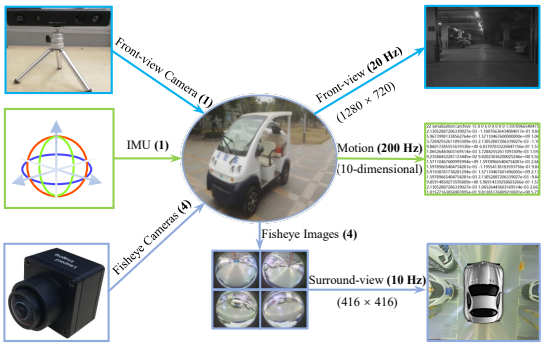

Figure 5: The configuration consists of a front-view camera, an IMU and four fisheye cameras forming a surround-view camera system to provide a surround-view image.

We perform parking-slots association in a strict manner, i.e.,

$$y_t^i = \begin{cases} k & f_t^i(k) \le th_1 \\ \varnothing & th_1 < f_t^i(k) < th_2 \\ n_t + 1 & f_t^i(k) \ge th_2, \end{cases} \tag{20}$$

where th_1 and th_2 are association and creation thresholds, which are empirically set based on the statistics of parking-slots' sizes. Specifically, when $f_t^i(k)$ is within the association threshold th_1, the observed parking-slot is associated with the k-th parking-slot in the map. When it is larger than the predefined creation threshold th_2, it means there is no associated parking-slot in the map, and a new parking-slot with ID $n_t + 1$ is created in the map. Otherwise, the parking-slot observation will be discarded.

5.3.2 Parking-slot Update. Once a parking-slot has a new observation, we need to update the position $\mathbf{L}_{y_t^k}$ of the parking-slot by the following equation, i.e.,

$$\mathbf{L}_{y_t^k} = (\sum_{n=1}^{n_t} \lambda^{n_t-n+1}\mathbf{T}_{CW}\mathbf{O}_t^k + {}_w\mathbf{P}_{y_t^k})/(n_t+1), \tag{21}$$

where n_t is the total number of parking-slots at time t. $\lambda = 0.9$ is the decay parameter, which implies that measurement at time t is more reliable than that at time $t-1$. Besides, the width of the parking-slot can be similarly updated.

6 EXPERIMENTAL RESULTS

Experiment Setup and Benchmark Dataset. We evaluated the proposed VIS_{SLAM} in an indoor parking site by driving an electric vehicle equipped with a front-view camera, an IMU and a surround-view camera system consisting of four fisheye cameras (Refer to Fig. 5 for details).

In order to facilitate the study of autonomous indoor parking algorithms, we have established and released a large-scale benchmark dataset. The dataset provides synchronized front-view images and surround-view images at 20 Hz and 10 Hz, respectively, with IMU measurements at 200 Hz. It contains 40,000+ front-view images and 20,000+ surround-view images, each of which was synthesized from four fisheye images, covering a wide variety of real cases in indoor parking sites. 10-dimensional motion data between every two

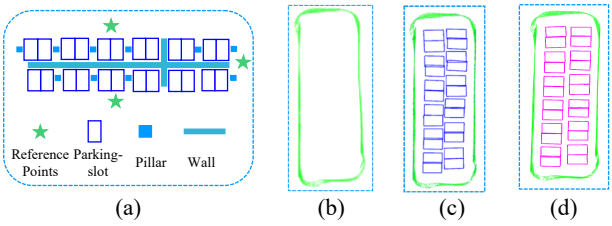

Figure 6: (a) A sketch of an indoor parking site from a top-down viewpoint. (b) Mapping result using visual and inertial error terms during optimization. (c) Mapping result by VIS_{SLAM} without a surround-view error term. (d) Mapping result by VIS_{SLAM} with a surround-view error term.

consecutive front-view images was also collected by IMU. The resolutions of the fisheye camera and the front-view camera are 1280 × 1080 and 1280 × 720, respectively. The spatial resolution of each surround-view image is 416 × 416, corresponding to a $10m \times 10m$ flat physical region, i.e., the length of 1 pixel in the surround-view image corresponds to $2.40cm$ on the physical ground.

Qualitative Results of VIS_{SLAM}. To qualitatively validate the effectiveness of the proposed VIS_{SLAM}, we drove the electric vehicle around an indoor parking site at around 10 km/h and then compared semantic maps using different error terms during optimization. Additional videos can be found in the supplementary material.

Fig. 7 (a) depicts a sketch of the indoor parking site from a top-down viewpoint. Fig. 6 (b) illustrates the result incorporating both a visual and an IMU error terms during optimization. It records the driving path and maps the 3D landmarks in the indoor parking site (3D landmarks are omitted here for display). However, parking-slots on the ground that are essential for autonomous indoor parking are not incorporated in the map. Fig. 6 (c) and Fig. 6 (d) demonstrate the results when the vehicle is equipped with a surround-view camera system, both of which construct not only 3D landmarks but parking-slots detected in surround-view images. Fig. 6 (c) shows the result without incorporating the surround-view error term during optimization. Since the scale estimated by IMU is difficult to be absolutely accurate, two rows of parking-slots are considerably approximate with each other, which violates the real situation where there is a wall between. In addition, there are obvious overlaps between adjacent parking-slots and a certain dislocation of the upper parking-slots. All above is due to accumulated errors caused by localization, surround-view calibration and parking-slot detection collectively. Fig. 6 (d) shows the result with the surround-view error term during optimization. The pink color denotes adjacency property between two parking-slots. When a surround-view error term is taken into consideration in optimization, the overall scale is more reasonable. The distance between each pair of adjacent parking-slots and the distance between two rows of parking-slots are more in line with the spatial distribution of the real scene. Besides, the overlapping area of each pair of adjacent parking-slots is significantly diminished, and parking-slots at the upper turning point are basically parallel with other parking-slots.

Comparison with Other SLAM Systems. Table 1 shows comparison of VIS_{SLAM} and eight existing representative SLAM systems from the viewpoint of three aspects, sensor modalities used ('S' for surround-view camera system), categories of map constructed, and whether parking-slots (PS) are incorporated in the map. It can be seen from the table that our VIS_{SLAM} is the first to incorporate surround-view camera system. It not only constructs semantic maps with parking-slots in the environment, but leverages no other information like Fiducial Tags used in [30] during optimization.

Table 1: Comparison with other methods.

Method	Sensors	Map	PS
Bowman *et al.* [2]	V (Visual)	Semantic	×
Civera *et al.* [3]	V	Semantic	×
Mur-Artal *et al.* [15]	V + I (IMU)	Geometric	×
Qin *et al.* [17]	V + I	Geometric	×
Tateno *et al.* [21]	V	Semantic	×
Yang *et al.* [25]	V	Semantic	×
Yu *et al.* [26]	V	Semantic	√
Zhao *et al.* [30]	V + I + T (Tag)	Semantic	√
VIS_{SLAM}	**V + I + S**	**Semantic**	√

Table 2: Revisiting errors of selected test points. (unit: meter)

Round	X	Y	Z	ΔX	ΔY	ΔZ	ΔD
Point 1 (-3.60 -0.80 15.73)							
Rd. 1	-3.61	-0.83	15.77	0.01	0.03	-0.04	0.051
Rd. 2	-3.60	-0.82	15.70	0	0.02	0.03	0.037
Rd. 3	-3.62	-0.83	15.75	0.02	0.03	-0.02	0.041
Point 2 (-16.79 -1.78 35.05)							
Rd. 1	-16.73	−1.77	35.08	-0.06	-0.01	-0.03	0.068
Rd. 2	-16.78	-1.77	35.07	-0.01	-0.01	-0.02	0.024
Rd. 3	-16.84	-1.77	35.03	0.05	-0.01	0.02	0.055
Point 3 (-17.07 -0.45 10.2)							
Rd. 1	-17.06	-0.45	10.25	-0.01	0	0.02	0.022
Rd. 2	-17.07	-0.45	10.31	0	0	-0.04	0.04
Rd. 3	-17.06	-0.45	10.28	-0.01	0	-0.01	0.014

Revisiting Error. Since it is difficult to obtain the ground truth of driving path, we can evaluate the localization accuracy by measuring the "revisiting error". Revisiting error is valid in localization evaluation in SLAM system because an autonomous parking system allows for an absolute localization error during navigation. As long as the revisiting error is small enough, the vehicle will adopt a consistent driving strategy when it drives to the same position.

In actual operation, the driver first manually drove the vehicle at around 10 km/h and the map was then initialized. Three map points at different locations were selected as reference points for test (Refer to Fig. 6(a)). Specifically, we chose two at the midpoints of both sides of the indoor parking site and one at the corner. After the map was stabilized (usually the vehicle should be driven for about three rounds), we evaluated by manually driving the vehicle to revisit three selected reference points, and recording the current coordinates at the test points. Then the differences in X-direction, Y-direction and Z-direction between the test points and reference points can be obtained. The final revisiting errors ΔDs were computed by adding up errors in all directions. Revisiting errors on all three reference points (Point 1, Point 2, and Point 3) are presented

Table 3: Gaps of adjacent parking-slots w/o surround-view error terms. (unit: meter)

Parking-slot	1	2	3	4	5	6	7	8	9	10	11	12	Mean
Without surround-view error terms	0.80	0.11	0.32	0.23	0.098	0.029	0.21	0.24	0.20	0.27	0.25	0.19	0.246
With surround-view error terms	0.18	0.18	0.074	0.073	0.096	0.060	0.11	0.08	0.077	0.06	0.073	0.21	0.106
Improvement	0.62	-0.07	0.246	0.157	0.002	-0.031	0.10	0.16	0.123	0.21	0.177	-0.02	0.140

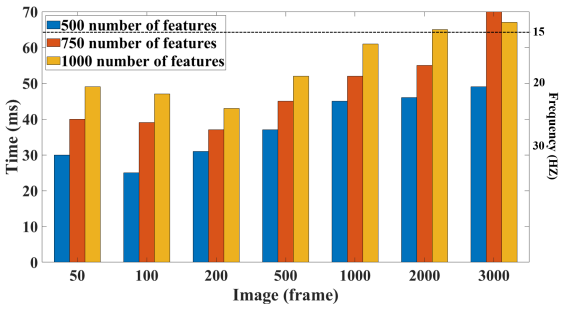

Figure 7: Average processing time per frame using different number of features.

in Table 2. It can be seen from Table 2 that the revisiting error of VIS_{SLAM} at each test point is less than 0.1m. Additionally, from Table 4, we can see that VIS_{SLAM} gains 64% of the favor compared with the revisiting error of 0.28m in [30], confirming the superiority of localization accuracy with VIS_{SLAM}.

Table 4: Comparison of revisiting errors with [30].

Methods	Zhao *et al.* [30]	**VIS_{SLAM}**
Average (unit:meter)	0.28 m	**0.08 m**

Distances of Adjacent Parking-slots. Since the adjacent parking-slots share a common marking-point, the gap between them is theoretically zero. By calculating the gaps of all groups of adjacent parking-slots, we can see from Table 3 that the averaged gap of adjacent parking-slots undergoes a dramatic decrease by 0.146m, a 57% decrease, if surround-view error terms are incorporated in optimization, which demonstrates the accuracy of the map constructed by VIS_{SLAM}.

Real-time Performance. We recorded the average processing time per frame of VIS_{SLAM} at running speed of 8-15 km/h. The result is presented in Fig. 7. It can be seen that when 1000 number of features are used, the average processing time per frame within 500 frames is 0.052s, that is, the frame rate can reach 20 fps. When the vehicle trajectory loops at around 3000 frames, the average processing time per frame is 0.067s, reaching 15 fps, which is qualified when driving in an indoor parking site at a low speed. In fact, the frame rate of the system can be improved by changing the number of extracted feature points. When the number of features extracted in VIS_{SLAM} is set as 500/750, the running speed undergoes a considerable improvement. Therefore, we can reduce the number of extracted feature points by sacrificing a certain degree of accuracy, if there is requirement for a higher frame rate.

Ablation Study. We performed detailed ablation analyses to validate the contribution of each error term during optimization of VIS_{SLAM} in three respects, the revisiting error, the average distance between adjacent parking-slots and the time cost, and the results are presented in Table 5. It can be seen from the table that both the revisiting error and the time cost of an visual-inertial error term based SLAM system can reach satisfied performance, which are 0.199m and 0.045s/frame, respectively. But it is not suitable for autonomous indoor parking, since the VI-SLAM system provides no semantic information of parking-slots during navigation. If we simply incorporate parking-slots in the surround-view image for tracking without optimization, they will compromise the SLAM system and lead to a huge revisiting error. But if the parking-slots are incorporated in optimization, both the revisiting errors and the adjacency gaps can be significantly diminished, confirming the effectiveness of VIS_{SLAM}. In addition, the time cost of VIS_{SLAM} is about 0.07s/frame (15 fps of the frame rate), which can be acceptable for an autonomous parking system running at a moderate speed.

Table 5: Optimization results using various error terms.

Mode	Revisiting error (m)	Adjacency gap (m)	T (s/frame)
V-I	0.199	-	0.045
S	0.317	0.243	0.040
VIS	0.028	0.106	0.067

7 CONCLUSION

In this paper, we proposed a tightly-coupled semantic SLAM system, VIS_{SLAM}, with configuration of a front-view camera, an IMU and a surround-view camera system mounted with four cameras around the vehicle. Parking-slots in the surround-view image are leveraged for optimization in order to improve the performance of VIS_{SLAM}. The qualitative mapping results of indoor parking sites and quantitative analyses on both localization accuracy and mapping precision demonstrate the effectiveness of our proposed VIS_{SLAM}. Actually, it has already been deployed on an electric car. A large-scale benchmark dataset consisting of synchronous multi-sensor data from typical indoor parking sites was also collected, providing a reasonable evaluation platform for autonomous indoor parking algorithms.

8 ACKNOLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61973235, Grant 61672380, Grant 61936014, and Grant 61972285, and in part by the Natural Science Foundation of Shanghai under Grant 19ZR1461300.

REFERENCES

[1] Simon Baker and Iain Matthews. 2004. Lucas-Kanade 20 Years On: A Unifying Framework. *International Journal of Computer Vision* 56, 3 (2004), Springer Nature Journal, 221–255. https://doi.org/10.1023/B:VISI.0000011205.11775.fd

[2] Sean L Bowman, Nikolay Atanasov, Kostas Daniilidis, and George J Pappas. 2017. Probabilistic Data Association for Semantic SLAM. *IEEE International Conference on Robotics and Automation (ICRA'17)* (2017), IEEE, Singapore, 1722–1729. https://doi.org/10.1109/ICRA.2017.7989203

[3] Javier Civera, Dorian Galvezlopez, Luis Riazuelo, Juan D Tardos, and Jose Maria Martinez Montiel. 2011. Towards Semantic SLAM Using a Monocular Camera. *IEEE International Conference on Intelligent Robots and Systems (IROS'11)* (2011), IEEE, California, USA, 1277–1284. https://doi.org/10.1109/IROS.2011.6048293

[4] Ronald Clark, Sen Wang, Hongkai Wen, Andrew Markham, and Niki Trigoni. 2017. VINet: Visual Inertial Odometry as a Sequence to Sequence Learning Problem. *arXiv:1701.08376* (2017).

[5] Hugh Durrant-whyte and Timothy S Bailey. 2006. Simultaneous Localization and Mapping: part I. *IEEE Robotics and Automation Magazine* 13, 2 (2006), 99–110. https://doi.org/10.1109/MRA.2006.1638022

[6] Paul Furgale, Joern Rehder, and Roland Siegwart. 2013. Unified Temporal and Spatial Calibration for Multi-sensor Systems. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'13)* (2013), IEEE/RSJ, Tokyo, Japan, 1280–1286. https://doi.org/10.1109/IROS.2013.6696514

[7] Dorian Galvezlopez and Juan Tardos. 2012. Bags of Binary Words for Fast Place Recognition in Image Sequences. *IEEE Trans. Robotics* 28, 5 (2012), 1188–1197. https://doi.org/10.1109/TRO.2012.2197158

[8] Chris Harris and Mike Stephens. 1988. A Combined Corner and Edge Detector. *Alvey vision Conference* (1988), Manchester, USA, 147–151. https://doi.org/10.5244/C.2.23

[9] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. 2015. Keyframe-based Visual-inertial Odometry Using Nonlinear Optimization. *International Journal of Robotics Research* 34, 3 (2015), 314–334. https://doi.org/10.1177/0278364914554813

[10] Mingyang Li and Anastasios Mourikis. 2012. Improving the Accuracy of EKF-based Visual-inertial Odometry. *IEEE International Conference on Robotics and Automation (ICRA'12)* (2012), IEEE, Minnesota, USA, 828–835. https://doi.org/10.1109/ICRA.2012.6225229

[11] Anastasios Mourikis and Stergios Roumeliotis. 2007. A Multi-state Constraint Kalman Filter for Vision-aided Inertial Navigation. *IEEE International Conference on Robotics and Automation (ICRA'07)* (2007), IEEE, Rome, Italy, 3565–3576. https://doi.org/10.1109/ROBOT.2007.364024

[12] Rodrigo Munguía, Emmanuel Nuno, Carlos I. Aldana, and Sarquis Urzua. 2016. A Visual-aided Inertial Navigation and Mapping System. *IEEE International Journal of Advanced Robotic Systems* 13, 3 (2016), 94–112. https://doi.org/10.5772/64011

[13] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan Tardos. 2015. ORB-SLAM: a Versatile and Accurate Monocular SLAM System. *IEEE Trans. Robotics* 31, 5 (2015), 1147–1163. https://doi.org/10.1109/TRO.2015.2463671

[14] Raul Mur-Artal and Juan Tardos. 2017. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robotics* 33, 5 (2017), 1255–1262. https://doi.org/10.1109/TRO.2017.2705103

[15] Raul Mur-Artal and Juan Tardos. 2017. Visual-Inertial Monocular SLAM With Map Reuse. *IEEE Robotics and Automation Letters* 2, 2 (2017), 796–803. https://doi.org/10.1109/LRA.2017.2653359

[16] Raul Murartal and Juan Tardos. 2014. Fast Relocalisation and Loop Closing in Keyframe-based SLAM. *IEEE International Conference on Robotics and Automation (ICRA'14)* (2014), IEEE, Hong Kong, 846–853. https://doi.org/10.1109/ICRA.2014.6906953

[17] Tong Qin, Peiliang Li, and Shaojie Shen. 2018. Vins-mono: A Robust and Versatile Monocular Visual-inertial State Estimator. *IEEE Trans. Robotics* 34, 4 (2018), 1004–1020. https://doi.org/10.1109/TRO.2018.2853729

[18] Tong Qin and Shaojie Shen. 2018. Online Temporal Calibration for Monocular Visual-Inertial Systems. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18)* (2018), IEEE/RSJ, Madrid, Spain, 3662–3669. https://doi.org/10.1109/IROS.2018.8593603

[19] Edward Rosten, Reid Porter, and Tom Drummond. 2010. Faster and Better: A Machine Learning Approach to Corner Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 1 (2010), 105–119. https://doi.org/10.1109/TPAMI.2008.275

[20] Ke Sun, Kartik Mohta, Bernd Pfrommer, Michael Watterson, Sikang Liu, Yash Mulgaonkar, Camillo Taylor, and Vijay Kumar. 2018. Robust Stereo Visual Inertial Odometry for Fast Autonomous Flight. *IEEE Robotics and Automation Letters* 3, 2 (2018), 965–972. https://doi.org/10.1109/LRA.2018.2793349

[21] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. 2017. CNN-SLAM: Real-time Dense Monocular SLAM with Learned Depth Prediction. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)* (2017), IEEE, Puerto Rico, USA, 6565–6574. https://doi.org/10.1109/CVPR.2017.695

[22] Massaki Wada, Kang Sup Yoon, and Hideki Hashimoto. 2003. Development of Advanced Parking Assistance System. *IEEE Trans. Industrial Electronics* 50, 1 (2003), 4–17. https://doi.org/10.1109/TIE.2002.807690

[23] Stephan Weiss and Roland Siegwart. 2011. Real-time Metric State Estimation for Modular Vision-inertial Systems. *IEEE International Conference on Robotics and Automation (ICRA'11)* (2011), IEEE, Shanghai, China, 4531–4537. https://doi.org/10.1109/ICRA.2011.5979982

[24] Oliver J. Woodman. 2017. An Introduction to Inertial Navigation. *https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-696.pdf* (2017), 1222–1229.

[25] Shichao Yang, Yu Song, Michael Kaess, and Sebastian Scherer. 2016. Pop-up SLAM: Semantic Monocular Plane SLAM for Low-texture Environments. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'16)* (2016), IEEE/RSJ, Daejeon, Korea, 1222–1229. https://doi.org/10.1109/IROS.2016.7759204

[26] Chao Yu, Zuxin Liu, Xinjun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei. 2018. DS-SLAM: A Semantic Visual SLAM Towards Dynamic Environments. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18)* (2018), IEEE/RSJ, Madrid, Spain, 1168–1174. https://doi.org/10.1109/IROS.2018.8593691

[27] Lin Zhang, Junhao Huang, Xiyuan Li, and Lu Xiong. 2018. Vision-based Parking-Slot Detection: A DCNN-based Approach and a Large-scale Benchmark Dataset. *IEEE Trans. Image Processing* 27, 11 (2018), 5350–5364. https://doi.org/10.1109/TIP.2018.2857407

[28] Zhengyou Zhang. 2000. A Flexible New Technique for Camera Calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 11 (2000), 1330–1334. https://doi.org/10.1109/34.888718

[29] Zhe Zhang, Shaoshan Liu, Grace Tsai, Hongbing Hu, Chen Chi Chu, and Feng Zheng. 2017. PIRVS: An Advanced Visual-inertial SLAM System with Flexible Sensor Fusion and Hardware Co-design. *arXiv:1710.00893* (2017).

[30] Junqiao Zhao, Yewei Huang, Xudong He, Shaoming Zhang, Chen Ye, Tiantian Feng, and Lu Xiong. 2019. Visual Semantic Landmark-Based Robust Mapping and Localization for Autonomous Indoor Parking. *Sensors* 19, 1 (2019), 161–180. https://doi.org/10.3390/s19010161