Vision based Semantic Mapping and Localization for Autonomous Indoor Parking

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

***Abstract*—Autonomous indoor parking without human intervening is one of the most demanded and challenging tasks of autonomous driving system. The key point to this task is real-time precise indoor localization. However, most indoor parking lots are composed of monotonous texture-less scenes and thus, are hostile to traditional visual feature-based SLAM methods. In this paper, we proposed a novel and practical solution of real-time indoor localization for autonomous driving in parking lots. High level landmarks, the parking slots, are extracted and enriched with labels to avoid the instability of low-level visual features. We then proposed a robust method for detecting incorrect data associations between parking slots and further extended traditional optimization framework by dynamically eliminating suboptimal data associations. Visual fiducial Tags are also introduced to improve the overall precision. Their number and distribution are also analyzed and compared. As a result, a semantic map of parking lot can be established fully automatically and robustly. We experimented the performance of real-time localization based on the map using our autonomous driving platform TiEV and the average accuracy of 0.3m tracking tracing can be achieve at a speed of 10kph.**

***Index Terms*—Indoor, ParkingLot, Semantic landmark, Robust SLAM**

I. Introduction

AUTONOMOUS driving has been witnessed great progress in recent years, breakthrough has been made in several harsh fields, including obstacle detection, real-time motion planning and high precision localization (mostly based on differential GNSS). Recently, testing self-driving car can already drive safely in urban and suburban areas[**?**]. However, parking in a large indoor parking lot without human interfere is still an unsolved problem. One critical reason is the lack of robust high precision localization mean in these GNSS forbidden area. Traditional indoor localization methods require pre-equipment sensors, such as WiFi, Bluetooth or UWB. Wireless signal suffers from occlusion and decays while users distance to signal sources increases, so a significant number of stations are needed for stability, let alone their relative low precision[**?**]. Laser-based SLAM system is eligible for localization an unmanned vehicle in environments such as a factory or a warehouse[**?**]. However, these range based representation is of high data volume and is vulnerable to dynamic scenes. As a result, visual SLAM (VSLAM) built on low-cost cameras became one of the most favorable localization method.

M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

Manuscript received April 19, 2005; revised August 26, 2015.

VSLAM is known to be effective in texture-rich environment[**?**]. Nevertheless, they can easily fail in monotonously textured scene such an indoor parking lot. **?**] adopted sparse feature point based SLAM method with panorama images to localize a car in parking lots. But the extracted sparse feature can be unstable when the ground floor is stained with tire markings or water spots. The distortion presented in the stitched panorama images can also disturb the feature extraction. Recently, **?**] employ DSO method with forward looking camera for mapping and localization in indoor parking lot. The direct methods estimate camera poses directly based on photometric error derived from the whole image, thus are more robust than sparse methods in less-textured area [**?**] [**?**]. However, they often require high frame rate and are susceptible to global illumination change, which restrict their usage in unevenly illuminated indoor parking lot[**?**].

Moreover, the performances of direct SLAM systems also depend on a good map initialization.[**?**][**?**] Most importantly, the re-localization based on pre-built dense map is not trivial since most direct methods are more like visual odometries rather than slam systems[**?**]. Yewei:In LSD SLAM, it is said that most dense method is simply a VO, while LSD has loop closure procedure. However, LSD itself do not have a map read and write module, but the LSD paper and the survey(Younes2016A) didn't mention this. Also since the direct method is not robust to illumination changes, re-localization using a pre-built map is of course almost unrealistic. Therefore, more stable and legible visual landmarks which are immune to various illuminate condition are demanded.

As a typical kind of semantic landmarks in parking lots, parking-space is now a favour for researchers [**?**] [**?**] [**?**]. Recently, deep learning-based method show its capability of accurate and robust detection of such kind of meaningful objects [**?**]. Inspired by these methods, we present an robust VSLAM system based on the recognition of high level landmarks for parking, i.e. parking-spaces and their IDs. Limited visual fiducial markers are introduced for improving overall accuracy and robustness. Facing the visual aliasing problem of parking slots, we proposed a robust outliers detection and elimination strategy in the optimization stage. Finally, a two dimensional map of parking slots can be robustly established which is distinguished from the traditional feature-based or point-cloud map for its stability, re-usability, light weight and human readable. Our system is implemented on an autonomous driving vehicle and tested in real parking lots.

II. RELATED WORKS

SLAM has long been a classic topic in the robotics field[**?**] and recently became heated in the autonomous driving since there are areas where GNSS is not available.[**?**]. **?**] **?**] and **?**] use Extended Kalman Filter(EKF) to simultaneously optimize the sensor and landmarks positions in real-time. These methods modeled the optimization problem as a Markov Chain, thus can provided promising localization results with high efficiency. However, the computation grows quadratically with the number of landmarks[**?**], so these methods are bounded in room-sized domains when regarding visual features as landmarks. Probability filters have many further extensions such as UKF [**?**], Information filter[**?**], particle filter[**?**][**?**] etc. However, they all assume the conditional independence of the current measurement with the historical states, which restricts the SLAM into a predict-update iteration loop. Inspired by the bundle adjustment research[**?**], a factor graph based optimization framework (known as Graph SLAM) was proposed [**?**]. The graph-based optimization method is closely related to a Markov random field model, thus can involve the influence of all historical measurements, which together with its flexibility enable the Graph SLAM became the most popular SLAM method[**?**].

Generally VSLAM methods fall into two groups, so called feature-based methods (the indirect methods) and direct methods. In feature-based methods, low-level features are detected and treated as landmarks. To emancipate tracking from the map-making procedure probabilistically, PTAM[**?**] separates tracking and mapping into two sub-tasks. Keyframes are extracted and used in optimization, but the system cannot deal with large loop closures, it can only handle tasks in small areas(a desk or a room corner). [**?**] and [**?**] build city-scale sparse maps using G2O, a graph optimization framework[**?**]. Nevertheless, the optimization of sparse 3D point cloud is time-consuming, so the off-line mapping procedure is a must. Built on the idea of PTAM[**?**], ORB-SLAM [**?**] offers stable and efficient graph-based VSLAM system. With the keyframe detection and the BoW-empowered fast loop closure detection, ORB-SLAM performs well in various indoor and outdoor environments. However, ORB-SLAM is still easy to fail in texture-less environments.

Because feature-based methods are only capable of creating sparse feature-based mapping, they cannot be directly used in applications where full reconstruction is demanded, e.g. AR or structure from motion. Direct methods based on photometric error and utilize all image pixels are proposed [**?**]. Comparing to sparse feature-based methods, direct methods can output a semi-dense point cloud with higher quality at a real time speed. But in practice, direct methods require high rate of overlapping between consequent frames, and high frame rate is also necessity since brightness consistency is crucial to estimate the depth accurately. Direct methods are also known to be vulnerable to motion blur, camera defocus and global illumination changes [**?**]. SVO [**?**] and DSVO [] combine advantages of feature-based method and direct method, and runs extremely fast (100 Hz). However, lacking of loop closure detection, these odometric methods drifts as time increases and gets lost easily.

Traditional SLAM methods do not incorporate human understandable meanings (semantics) associated with landmarks into the method, which now is recognized to be crucial for construct a human readable map and strengthen the descriptive power of landmarks[]. [**?**]. [**?**] added semantic labels to a LSD-SLAM framework to construct a dense map with classes attached to geometric entities, but semantic labels helps little in the optimization or localization stages. SLAM++ [**?**] and Semantic Fusion [**?**] employed semantic labels in the RGBD SLAM framework to aid the loop closure. However, both methods...

JOHN!!should be very clear about the defects of these methods, i.e. 3D objects? not precise regarding to the localization of objects? can be confusing when multiple objects are searched?

In [**?**], shop names and shop facades are recognized as labels in large indoor shopping spaces. [**?**] detects parking lots from the (fixed IP camera?) top view image, and use detected lots to estimate camera poses (is this a slam?).

[**?**] and [**?**] reconstruct the metric map and the semantic map of parking lot, which helps the route planning and parking task. (!!semantics in VW paper(2015) is not used in SLAM at all, not sure about the second one)

In a short conclusion, existing VSLAM methods generally could not perform robustly in texture-less area like an indoor parking lot. Therefore, more descriptive landmarks, especially landmarks attached with semantics should be used instead of low level feature in such a scenario. JOHN add review of robust method or include this in the approach?

YEWEI: what is considered a robust method?Max-mixture...

III. APPROACH

Our semantic VSLAM system includes four fisheye cameras and one monocular camera. Four fisheye cameras are fixed at two reflectors, and at the front and rear bumpers, which consist a surround-view system. A top-view image is then fused from the surround-view inputs after intrinsic and extrinsic calibration, as shown in Fig. 2. In the top-view image, who indicates ground textures, parking slots are detected. The monocular camera is installed to the left of the real-view mirror to capture front scenes. The change of steering wheel angle, as well as the vehicle speed and heading direction collected by IMU are also used in our system.

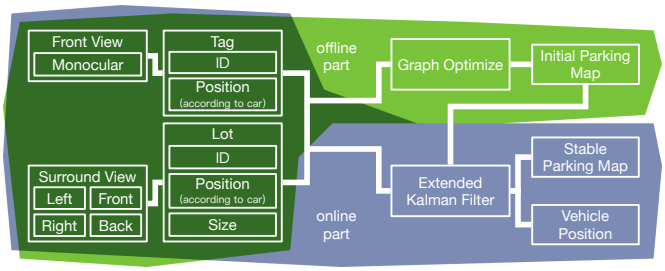

Fig. 1. Pipeline of the method

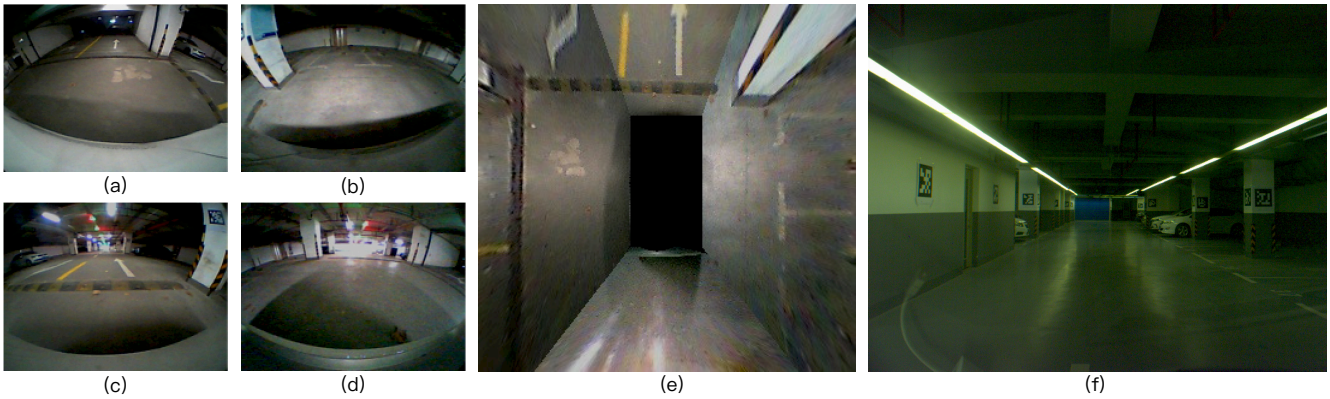

Fig. 2. (a)(b)(c)(d) are images from left, right, front, back fisheye cameras. (e) shows the top-view image fused from (a)(b)(c)(d). The image from the monocular camera is (f).

We choose parking slot as the landmark since parking slots are the most distinguishable objects in a parking lot, and the precise locations of parking slots are informative for localization and navigation during autonomous parking. Our parking slot detector is based on [**?**], in which corner points of parking slots are detected (Fig. **??**).

Yewei:Parking slots are not classified until the parking slots hypothesis check, and the "T" type is quite hard to be classified to pattern (a) or (b) in Fig.7 of [**?**] Fig. 5 in [**?**] is to illustrate the training examples of corner detection procedure.

Afterwards, parking slots are assembled according to the image patterns around corner points (Fig. **??**). Although the CNN-based method is capable of detection most kinds of corner points fast and robust, the exact shape of the parking slot cannot be known due to the limited visible range of the surround vision system. As a result, the parking slot can only be guess initially and we have to optimize the shape of the parking slots in the SLAM system. Furthermore, the ID of each parking slot should be detected for facilitating data association between parking slots, which will be elaborated in

Another kind of landmark used in our system is visual fiducial markers. Fiducial markers are introduced as an aid for the constancy of localization since few parking lot is detected near the entrances and exits. In the practice, we find that loop closure using only parking slots is not robust enough, so visual markers are also placed where there is a high revisiting-rate. Fiducial markers may not be fully prohibited but their number can be limited to a small amount. We select AprilTags as visual markers for its robustness and high efficiency(a frame rate of 10 Hz)[**?**].

A. CNN based Parking slot Recognition

We adopt the method proposed by [**?**] to detect parking slots.

John: Donot repeat existing method, briefly describe the detection method and emphasize on the post processing if there is any and the ID detection

The raw detection results from the previous method are still not precise. We remove the entrance-line candidate who has more than two marking-point patterns on it to avoid this situation. Extremely large or small candidates are also discarded as all slots are around the same size. Once a parking-line is detected, the direction of a lot is determined as well. And the depth of a parking lot is known a prior knowledge.

Since the relative car coordinate overlaps the top-view image coordinate, the slot coordinate in the top-view image is the same as that in the vehicle relative coordinate, and thus, we only need to perform a scale transformation. To optimize a slot's shape as well as its coordinate, a slot is stored as four landmarks other than one. Each landmark is connected with other three ones with a rectangular constrain(Fig. **??**??)), which consists of an angular constraint of high confidence and a distance constraint with a relatively lower confidence. The observation between four sub-landmarks of a slot and vehicle position is also applied to connect the slot to the global map.

ID of a parking slot is important for the association of this semantic landmark. We fine tuned PVANet to detect each character in one ID [**?**]. The entrance-line of the parking slot is firstly utilized to roughly locate the ID, as shown in Fig. **??**??). Then the image patches are extracted and send to the CNN (Fig. **??**??)). Due to the distorted and blurred texture in the surround view image, even the latest detection network could not offer the satisfactory performance. As a result, we devised a semantic-based fuzzy association method to cope with the uncertainty, which will be detailed in **??**.

B. Visual fiducial marker localization

The surround view system offers an intuitive ground observation at a high frame rate, however it has limitations. The visibility range and resolution of the top-view image is far from satisfactory, limiting the slot detection performance. And the calibration for image fusion becomes inaccurate as time goes, which will deteriorate the slot measurement.

Recalling the goal of our robustly localization system for autonomous driving and parking in various parking lots, certain numbers of faithful landmarks such as visual fiducial markers still have to be incorporated.

We adopted AprilTag and employed the detection framework from the AprilTag C source Open Library [1][**?**]. We

[1]https://april.eecs.umich.edu/wiki/AprilTags

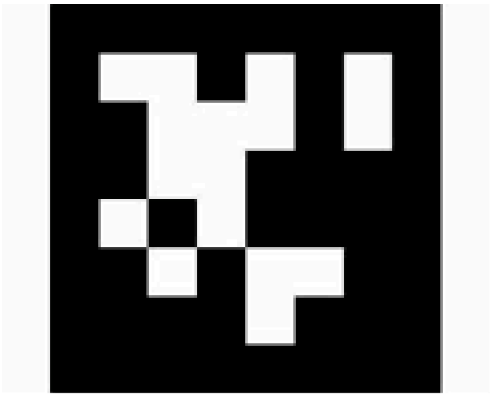

Fig. 3. An example of AprilTag

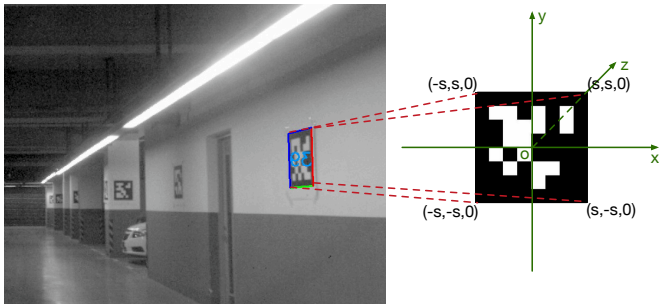

Fig. 4. An example of the relationship between a detected tag in the image plane(left figure) and its corresponding hypothetical 3D tag coordinate(right figure)

further solved the relative position between visual markers and the vehicle by the PnP model in a fast and accurate way [**?**].

John: Can we improve the efficient or the accuracy when localization a tag?

In the PnP model, we assume four corner points of the tag always lie on the same plane, and thus definite the hypothetical 3D tag coordinate(whose origin is a tags geometrical center, x axis and y axis are both parallel to tags edge, pointing rightward and upward respectively, and z axis is perpendicular to the tag plane, pointing inward). So the relationship between the tag corner in the image and that in the hypothetical 3D tag coordinate can be described as, $R_{3\times3} \cdot {K_{3\times3}}^{-1} \cdot x_{3\times1} + t_{3\times1} = X_{3\times1}$ where $R_{3\times3}$ and $t_{3\times1}$ is the relative rotation and translation vector, $K_{3\times3}$ is the intrinsic camera matrix, $X_{3\times1}$ is the coordinate in the hypothetical 3D tag coordinate and $x_{3\times1}$ is the homogeneous coordinate in the image plane. $X_{3\times1}$ and $K_{3\times3}$ are prior knowledge, and $x_{3\times1}$ has already been detected. Only $R_{3\times3}$ and $t_{3\times1}$ are unknown parameters with 6 DOF, more than 3 corresponding points are needed for iteration.

In experiment, we find $R_{3\times3}$ given by PnP method cannot always meet our demands in accuracy. So we discard $R_{3\times3}$ and get angle directly from the calibrated image. As shown in(Fig. **??**??)), the angle(α) can simply be calculated by $\alpha = \arctan((x_i - x_0)/f)$, where x_i and x_0 denote the x coordinate value of the tag centre and the principle point respectively, and f is the focal length. The distance(d) is the 2-norm of $t_{3\times1}$. So the tag locates at $(x = sin(a) \cdot d, y = cos(a) \cdot d)$ in the vehicle relative coordinate.

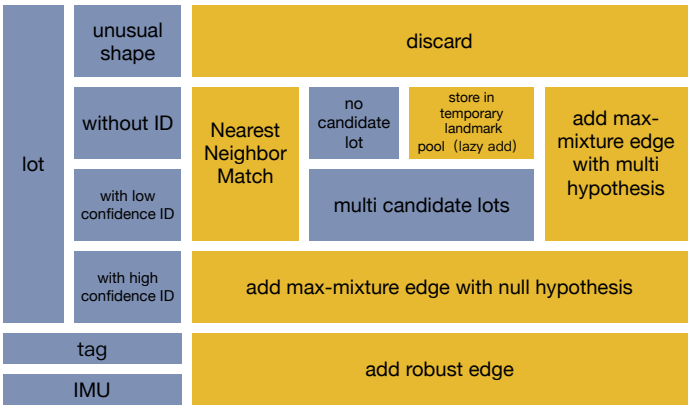

Fig. 5. The overall pipeline of optimization

These visual fiducial markers are flexible and easily implemented. They brought another benefit for the autonomous parking purpose, that fiducial markers can easily indicate the existences of pillars and walls which can only be robustly detected by expensive laser scanners. These obstacle information can facilitate the route planning inside of a parking lot.

C. Optimization

1) Optimization Pipeline: False positives occur in both detection and data association of parking slots, and affect map and localization result greatly. In front-end part, several methods have successfully lowered the slot detection error, while wrongly associated slots still cannot be detected. Thus, error detection and fixation in slot detection and association should be performed in both front-end and back-end.

At every frame, the relative car-landmark position(including the slots and visual markers), together with the speed and angular changes of vehicle are collected to the parking map incrementally. These values are achieved from a Kalman-based extrapolator with the steering wheel, car speed and the compass readings from a cheap IMU as the inputs. Parking slot observations are pre-associated roughly through their ids and the nearest neighbour search. However, due to fallible detection of parking slots and their ids from low-quality surround vision images, ambiguities will be presented during data association. Max-Mixture is therefore introduced and improved to not only detect but also correct mistaken associations. And the overall of the optimization pipeline is shown in Fig. 5.

2) Optimization Framework: [**?**] offers a global framework for nonlinear least square problems. Under the G2O framework, all the elements to be optimized are regarded as vertices, while all the observation constraints are known as edges who connect vertices. In our method all the observations of parking slots and fiducial markers are treated as edges in the graph. However, ambiguity raised from observations of parking slots as well as parking ids lead to erroneous data associations and generate fatal residual error in the optimization.

Several robust optimization extensions exist, in the paper, max mixture models[**?**] are introduced to detect and eliminate the mistaken associations. The original least square problem of graph optimization can be expressed by the following equation:

John: abbreviate the explanation of the g2o, but should add the robust method and emphasize on our contributions

$$F(x) = \sum_{k \in C} e_k(x_k, z_k)^T \Omega_k e_k(x_k, z_k) \text{[?]}$$

$$x^* = \underset{x}{argmin} F(x) \text{[?]}$$

where C denotes the total ID group; $x_k$is the parameter block of the kTh vertex(containing the vertexs location and direction), z_k and Ω_k is the mean and covariance of the kTh edge; $e_k(x_k, z_k)$ measures how well x_k matches z_k. These elements consist the global error function $F(x)$, and x^* is the global optimal solution.

We can get x^* by solving the linear function iteratively, and the uncertainties of z_k are described by noise models who are uni-model Gaussian. This classical graph framework works quite well when there are only visual markers and IMU measurements in parking map, but it gives bad results once parking slots are added. Since every parking slots may be wrongly associated or detected, it's impractical to measure their noises by a single model.

Yewei: highlight, comparing to the traditional methods, why semantic, how outliers

3) Outliers elimination using Max-Mixture Model: As is mentioned above, classical graph optimization method using uni-model Gaussian is sensitive to outliers, and fails when there are wrongly associated datum in graphs. Several robust methods**?**] **?**] **?**] have been proposed for pose graph, but both poses and landmark positions are to optimize in our system.

Max-Mixture model share some similarities with slot ambiguity, so it can help fix data association problem and to some extent, implement semantic information. Observation noises are described by multi-model Gaussian, thus wrongly associated data can be suppressed by other mixture elements. To simplify the problem, the sum is substituted by a max operator. In the pose graph situation, when an uncertain loop closure occurs, a Max-Mixture of several elements is thus added to the graph: several possible loop closure alternatives and a null hypothesis indicating that all the former alternatives are wrong. In the parking map with loop closures triggered by observing a pre-visited landmark, Max-Mixture still works.

Only noises of parking slots are treated as multi-model Gaussian, while in other situations, uni-model Gaussian is enough. After the elimination based on prior knowledge in front-end, slots are classified by the confidence level of lot ID detections. Lot observations with high confidence IDs are added to the graph as Max-Mixture edges with null hypothesis. A null hypothesis is always essential due to the existence of false positive ID detection results with high confidence. The rest slots get several data association hypothesis by processing nearest neighbour search in the existing parking map. These candidate hypothesis together with the original hypothesis obtained by lot detection result consist a Max-Mixture landmark observation element with multi hypothesis. This element is then added to the graph, only one hypothesis will be selected during optimization. A new landmark hypothesis will not be added to the map until it is repeatedly observed for a certain number of time. When slot ID detection only achieves partially success(only one in two digits is detected), Max-Mixture can also help decide the full slot ID. Inspired by [**?**], information matrix of landmark edges are shifted according to their detection confidence level.

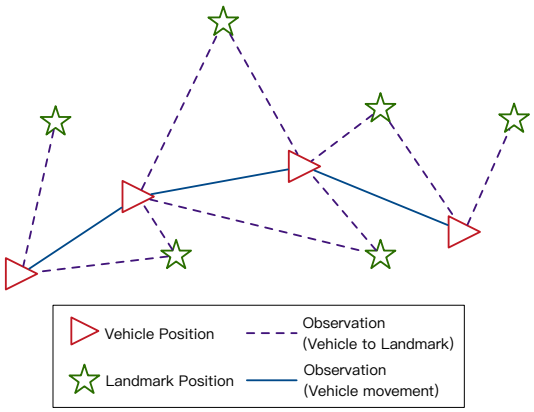

Fig. 6. The connection between the vehicle at each time stamp and the landmarks.

IV. Experiments

JOHN: should firstly describe the test environment, size map lighting condition and car etc. How outliers detection successful?

In this section, we test our method both online and offline, and discuss several questions about the usage of visual markers. All the parking lot datasets are collected and tested by TiEV autonomous vehicle [2] .

A. Semantic landmark detection and localization

John: models used, training workflow, training samples and detection results and statistics, short conclusion Fig. **??** show a subset of the training samples used.

John: offline should be placed before online

B. Offline mapping test

As to the offline part, several datasets with different starting points are collected. Each dataset is processed independently. The covariance of each edge shifts from 0.1 to 0.25 according to its confidence level. Fig 12 shows the map result for two of these datasets. Map blocks with more tags have higher precision as tags are ideal landmarks who will never be wrongly associated. The low detection rate of parking ID also affects the mapping result, as shown in fig 13.

In one dataset, there is a no-tag map block (marked in fig 12), where wrongly associated data always occurs. However, once a loop is detected, all landmark-positions are corrected. Then, duplicate landmarks are merged.

[2]cs1.tongji.edu.cn/tiev

Fig. 7. toDo

Fig. 8. toDo

C. Online real-time mapping and localization

During the online part, the vehicle is first operated by human to initialize the parking map. Once the map stabilizes, a car trace is recorded. Then the vehicle drives automatically at the speed of 3 km/h following this pre-recorded trace and fixes its direction constantly according to real-time localization record. The automatic driving procedure is repeated 10 times, automatic driving traces are also recorded and compared to the pre-recorded one. Fig 11 shows the traces of both manual and automatic driving trace. Quite a number of visual markers(60 tags) are used to cover the lot-free parts near the entrance and to guarantee a fully stable and credible localization result. Each tag is printed on a A2-size paper with 48.8 cm side length. While observing, those tags who are 20 meters or farther than the vehicle are discarded since the accuracy decreases as tags become smaller in the image. However, observations with dip-angle are reliable constraints as long as they are within 20 meters.

Fig. 9. The dip-angle doesn't affect the accuracy while the distance makes great influence.

Fig. 10. A comparison between the human-driving and automatic driving trace

D. Comparison with ORB2-SLAM

E. How many tag is needed?

As described in the previous section, tags are ideal landmarks promising a robust localization and mapping result. In this part, we discuss which tags are indispensable while others are not a must.

We think several factors may affect the importance of each tag: the visibility rate, the distribution and the distance to the nearest rate. Considering the factors above, we make several tag combinations and test their performance. The final result is shown in fig 14, both visibility and distribution play an important role. The crucial tags should be separated in the lot, and have a good visibility rate. In our parking lot case, at least 3 tags are needed, while 10 evenly distributed tags offer the best result.

a statistic result of the rightly located lots in each graph give the tag position solution based on the visibility of cameras based on visibility of the front camera(put tags at where there are multi visibility) be separated in the lot compare the solution to the statistic result very similar

John: should add Tags can be replaced by instruction arrows or parkinglot ids on the pillar, here restricted by the goal and the environment tags are required

V. Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgment

The authors would like to thank...

Fig. 11. toDo