2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Visual Semantic Localization based on HD Map for Autonomous Vehicles in Urban Scenarios

Huayou Wang*, Changliang Xue*, Yanxing Zhou, Feng Wen and Hongbo Zhang

***Abstract*— Highly accurate and robust localization ability is of great importance for autonomous vehicles (AVs) in urban scenarios. Traditional vision-based methods suffer from lost due to illumination, weather, viewing and appearance changes. In this paper we propose a novel visual semantic localization algorithm based on HD map and semantic features which are compact in representation. Semantic features are widely appeared on urban roads, and are robust to illumination, weather, viewing and appearance changes. The repeated structures, missed detections and false detections make data association (DA) highly ambiguous. To this end, a robust DA method considering local structural consistency, global pattern consistency and temporal consistency is performed. Further, we introduce a sliding window factor graph optimization framework to fuse association and odometry measurements without the requirements of high-precision absolute height information for map features.**

We evaluate the proposed localization framework on both simulated and real urban road. The experiments show that the proposed approach is able to achieve highly accurate localization with a mean longitudinal error of 0.43m, a mean lateral error of 0.12m and a mean yaw angle error of 0.11°.

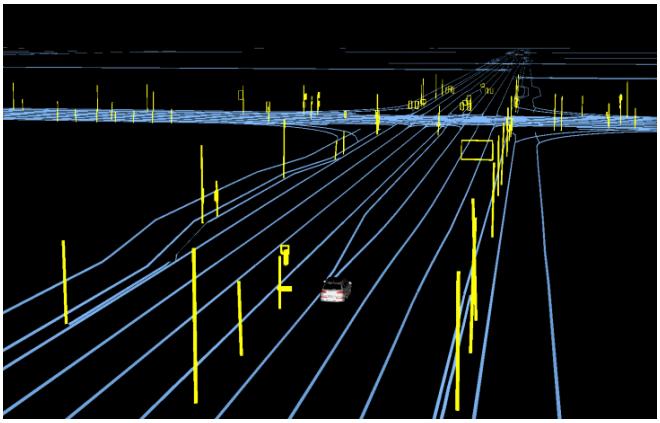


Fig. 1. Snapshot of evaluation drive with localization pose and HD map. The HD map consists of semantic features such as road markings, traffic lights, signs and poles.

I. INTRODUCTION

AVs have received widespread attention from industry and academia in recent years. Highly accurate localization is an essential technology for AVs, because various modules, such as decision-making, planning and control, are heavily dependent on positioning. To achieve accurate localization, AVs are equipped with various sensors, such as GNSS, camera, LiDAR, IMU, wheel encoder, etc. Due to the expensive price of LiDAR, low-cost camera and IMU are more suitable for localization of commercial-level AVs.

There are various complex road conditions, such as urban canyons, tunnels, viaducts, etc, in urban scene, which makes it more challenging for AVs. To achieve robust localization in this scene, various methods appeared, such as GNSS-based methods [1], [2], vision-based methods [3], [4], visual-inertial-based methods [5]–[8], LiDAR-based methods [9], [10]. GNSS-based method can achieve centimeter-level accuracy in open scene, but it is not reliable enough in occlusion and multipath conditions. The methods of fusing GNSS and IMU or odometry [11]–[15] are proposed to solve the problem of GNSS, but they still fail in scenes with long-term lack of global location information due to the drift of odometry. To solve the problem of drift, methods based on a priori map are widely applied. The most commonly used map is point cloud map which can achieve centimeter-level localization through ICP or NDT method, but the storage of point cloud map is a big challenge for commercial-level AVs. Traditional visual feature maps have also been tried for positioning, but they suffer from tracking lost due to illumination, weather, viewing and appearance changes.

To solve this problem, we extract visual semantic features and perform positioning based on HD map. Compared with traditional visual features, semantic features are widely appeared on urban roads and are long-term stable and robust to weather, illumination, viewing and appearance changes. DA is one of the biggest challenges due to the singularity, false and missed detections of semantic features. Therefore, we propose an accurate and robust visual semantic localization system with consistent DA. The main contribution of this paper is summarized as follows:

- An accurate and robust localization algorithm based on visual semantic features and lightweight HD map without the requirements of high-precision absolute height information for map features.
- A robust DA method based on local structural consistency, global pattern consistency and temporal consistency to resolve the ambiguity of DA.
- A factor graph optimization framework which tightly couples visual semantic measurements and odometry measurements for robust localization.
- Extensive experiments are carried out on both simulated and real urban roads to verificate the effectiveness of DA and accuracy of localization.

*Equal Contribution.

All authors are with Noah's Ark Lab, Huawei Technologies, Beijing, China. {wanghuayou, xuechangliang, zhouyanxing, wenfeng3,zhanghongbo888}@huawei.com.

II. RELATED WORK

A. Methods based on traditional visual features

Methods based on traditional visual features extract geometrical features, such as points, lines and planes, and perform feature matching through descriptors. Mul-Artal [3] and Sons [16] built feature map with ORB and BIRD descriptor respectively, then the pose can be obtained by feature matching. ETH ASL LAB has done extensive work in the field [17]–[21], including multisession map summary and appearance-based online landmark selection. However, these methods still cannot get rid of the influence of weather, illumination, viewing and appearance changes.

B. Methods based on semantic features on roads

Methods based on semantic features on roads are widely applied to AVs. Semantic features consist of road markings, traffic lights, traffic signs, poles, etc. Schreiber [22] and Poggenhans [23] detected road markings and curbs, and located AVs by matching features with map. Lu [24] applied chamfer matching to construct constraints of road markings and formulated a non-linear optimization problem to estimate the 6DoF pose. Moreover, Jeong [25] classified road markings to avoid ambiguity, and achieved precise localization through sub-map matching, loop closure and pose graph optimization. Wilbers [26] and Spangenberg [27] realized pose estimation through poles with depth. Meanwhile, Sefati [28] fused road markings and traffic signs from camera and LiDAR to perform positioning through PF. Further, Wu [29] performed localization through lane lines extracted from camera and blob features extracted from occupancy grid. Ma [30] fused INS, GPS, positioning results of lane lines and traffic signs through probabilistic histogram filtering. In addition to road markings and poles, Kummerle [31] also extracted the geometric information of the vertical surface of buildings through laser to realize precise localization.

In this paper, we only use a monocular camera for positioning. The work of Jin [32] and Xiao [33] is similar to ours, but we do not require the absolute height but the height relative to the road surface of current location. This simplifies the difficulty of map construction and greatly reduces costs.

C. Semantic data association

Due to singularity, false and missed detections of semantic features, it's extremely challenging to achieve correct and robust association. Spangenberg [27] associated perceived poles with map through Euclidean distance and width of poles. Meanwhile, Hu [34] and Xiao [33] applied RANSAC to eliminate mismatching. Further on, Kummerle [31] and Wilbers [26] constructed sub-map by accumulating detections over time to solve the ambiguities of DA. In the area of target-tracking, Hungarian algorithm [35] and Multiple Hypothesis tracking [36] were performed. To solve the singularity of DA, Bowman [37] introduced probabilistic DA into semantic SLAM system. In contrast to these methods, we proposed a robust DA method based on local structural consistency, global pattern consistency and temporal consistency to eliminate mismatching caused by singularity.

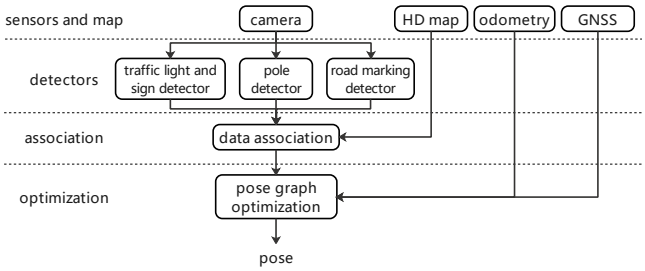


Fig. 2. Overview of the localization framework.

III. SYSTEM OVERVIEW

The global localization problem can be defined as: Given a serious of sensor measurements $\mathcal{Z} = \{\mathbf{z}_k\}_{k=1}^{K}$ and HD map $\mathcal{L} = \{\mathbf{l}_m\}_{m=1}^{M}$, estimate the pose sequence $\mathcal{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ which represents state trajectory. The pose $\mathbf{x}_t$ and feature position $\mathbf{l}_m$ are defined as $\mathbf{x}_t \in SE(2)$ and $\mathbf{l}_m \in \mathbb{R}^3$. The localization problem can be represented as the following maximum a posteriori (MAP) inference problem:

$$\hat{\mathcal{X}} = \arg\max_{\mathcal{X}} p(\mathcal{X}|\mathcal{Z}, \mathcal{L}) \tag{1}$$

The MAP problem can be divided into two steps, including DA process and pose estimation process base on DA. Equation (1) implies DA process, it is necessary to establish the DA $\mathcal{D} = \{\mathbf{d}_t\}_{t=1}^{T}$ between measurements and map based on prior pose $\mathcal{X}^0$ before performing pose estimation. Therefore, the MAP inference problem can be redefined as:

$$\begin{aligned} \hat{\mathcal{X}}, \hat{\mathcal{D}} &= \arg\max_{\mathcal{X},\mathcal{D}} p(\mathcal{X}, \mathcal{D}|\mathcal{Z}, \mathcal{L}) \\ &= \arg\max_{\mathcal{X},\mathcal{D}} p(\mathcal{X}|\mathcal{Z}, \mathcal{L}, \mathcal{D}) p(\mathcal{D}|\mathcal{X}^0, \mathcal{Z}, \mathcal{L}) \end{aligned} \tag{2}$$

Therefore, the localization framework is divided into four components, namely sensors and map, detectors, association and optimization, as shown in Fig. 2. Sensors consist of a monocular camera, an IMU, two wheel encoders and a GNSS receiver. The camera is used to detect semantic features. The IMU and wheel encoders form odometry to provide local relative motion estimation. The GNSS receiver can provide a rough estimation of current pose, which is used for system initialization. The detector layer detects road markings, poles, traffic lights and signs from image. The association layer associates the semantic features extracted from image to features in HD map. The association process is divided into five steps. First, generating proposals around prior pose and projecting map features into image based on each sampled pose. Second, coarse association based on local structure consistency is implemented to find an approximate optimal sampled pose. Third, an optimal association method considering matching number, matching similarity and local structure similarity is performed to achieve optimal global consensus matching. Then, feature tracking between consecutive frames is executed. Finally, temporal smoothing is performed to obtain temporal consistent DA. In the final optimization layer, pose graph optimization is introduced to estimate the pose based on DA and odometry measurements.

IV. METHODOLOGY

A. Semantic Features and Detection

The choice of semantic features is critical to localization performance. In this paper, we choose features based on the criteria proposed in reference [31], therefore road markings, poles, traffic lights and signs are selected for positioning. They are easy to detect, frequently appearing, memory efficient with compact representation, and weather, illumination, viewing and appearance invariant.

We adopt a popular Convolutional Neural Network (CNN) method YOLOV3 [38] to detect features. A detected sign $s_t = (s_t^l, s_t^c, s_t^b)$ comprises a detected class s_t^l, a score s_t^c denoting the detection confidence, and a bounding box s_t^b. The four contour points of signs are stored in HD map, and the height of each point is the height relative to the road surface of current location. A detected pole $s_t = (s_t^l, s_t^c, s_t^b)$ consists of a detected class s_t^l, a score s_t^c denoting the detection confidence, and s_t^b representing two vertices. Poles are stored in HD map with two vertices. Road markings are represented as sample points in image plane and HD map.

B. Semantic Data Association with HD map

Because of singularity, false and missed detections of semantic features, DA becomes one of the most challenging problems for semantic localization system. In this paper, we propose a robust DA method based on local structural, global pattern and temporal consistency to solve the ambiguity of DA and ensure spatial and temporal consistency. To illustrate the proposed DA method, the pseudocode is provided in Algorithm 1. The details of DA process is as follows:

Step 1: Generating proposals by sampling around prior pose obtained by odometry. For each sampled pose T_v, map features are projected into image plane:

$$p_i = \frac{1}{Z_i^c} K T_v^c T_v^{-1} P_i^m \tag{3}$$

where, P_i^m is the position of the ith map feature. K and T_v^c are intrinsic and extrinsic parameters of camera. Z_i^c is the z-axis position of the ith map feature in camera coordinate.

Step 2: Coarse association based on local structural consistency is performed to find approximate optimal sampled pose for eliminating mismatching caused by large prior pose error. Local structural consistency keeps the lateral position distribution of perceived features and corresponding reprojection features consistent. First, perceived and reprojection features are sorted in ascending order according to lateral position. Second, we calculate the similarity between each perceived feature s_t and each reprojection feature r_k:

$$p(s_t|r_k) = p(s_t^l|r_k^l)p(s_t^c|s_t^l, r_k^l)p(s_t^b|r_k^b) \tag{4}$$

where, $p(s_t^l|r_k^l)$ and $p(s_t^c|s_t^l, r_k^l)$ can be obtained by offline learning of perception results. For signs, the likelihood $p(s_t^b|r_k^b)$ consists of position and size similarity:

$$\begin{aligned} p(s_t^b|r_k^b) =& \omega \exp(-\frac{1}{2}(\frac{x_p - u_p}{\sigma_p})^2) + \\ & (1-\omega)\exp(-\frac{1}{2}(\frac{x_s - u_s}{\sigma_s})^2) \end{aligned} \tag{5}$$

Algorithm 1 Data association

Input: $\mathcal{X}^0$: initial pose; $\mathcal{L}$: HD map features; $\mathcal{Z}$: perceived features; $\mathcal{D}_{1:T-1}$: association results in sliding window;
Output: optimal $\mathcal{D}_T^*$

1: generate proposals around $\mathcal{X}^0$, and project $\mathcal{L}$ into image according to Eq. (3);
2: find best proposal according to Eq. (6);
3: construct multi-order graph matching problem, and compute $\mathcal{D}_T$ according to Eq. (7);
4: track features in consecutive frames;
5: perform temporal smoothing according to Eq. (9) to gain temporal consistent $\mathcal{D}_T$;
6: **return** $\mathcal{D}_T$

where, ω is a learned hyperparameter for weighting position similarity and size similarity. u_p, u_s, x_p and x_s denote the position and size of map feature and perceived feature respectively. σ_p and σ_s can be learned offline from perception results. Similar to signs, the likelihood $p(s_t^b|r_k^b)$ of poles consists of position, orientation and overlap similarity.

If the maximum similarity score of a perceived feature is larger than a threshold and the local structure is preserved, they are considered as a match pair. For each sampled pose, the cost C is computed to approximately evaluate it based on matching number N_m and matching error $e_{ii'}$:

$$C = -\frac{1}{N_m}\sum_{i=1}^{N_m} e_{ii'} + \omega N_m \tag{6}$$

where, ω is a hyperparameter. $e_{ii'}$ is defined as the lateral distance between feature i and i', as shown in Fig. 3. The proposal with max C is regarded as the approximate optimal matching sampled pose and will be used in step 3.

Step 3: Based on the approximate optimal matching sampled pose, an optimal association method considering matching number, matching similarity and local structure similarity is performed to achieve optimal global consensus matching. It is formulated as a multi-order graph matching problem by solving the following optimization problem:

$$\begin{aligned} \hat{\mathbf{X}} = \arg\max_{\mathbf{X}} & \ \omega_1 N_m + \omega_2 \frac{1}{N_m}\sum_{i=1}^{N}\sum_{i'=1}^{M} x_{ii'} s_{ii'} \\ & + \omega_3 \frac{1}{N_e}\sum_{i}^{N}\sum_{i'}^{M}\sum_{j}^{N}\sum_{j'}^{M} x_{ii'} x_{jj'} s_{ij,i'j'} \end{aligned} \tag{7}$$

S.t.

$$\sum_{i=1}^{N} x_{ii'} \le 1, \ \sum_{i'=1}^{M} x_{ii'} \le 1, \ x_{ii'} = 0 \ or \ 1$$

where, N and M are the number of perceived and reprojection features, N_e is the number of edges between two features. ω_1, ω_2 and ω_3 are hyperparameters. $x_{ii'}$ denotes if perceived feature i is matched with reprojection feature i'. $s_{ii'}$ represents the similarity between perceived feature i and

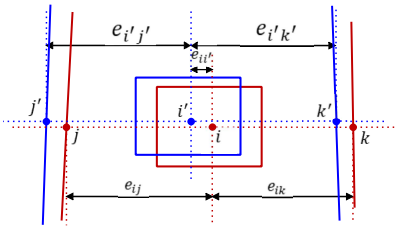


Fig. 3. Graph representation of matching. Perceived features (two poles placed on both sides of a sign) are denoted in red, reprojection features are expressed in blue.

reprojection feature i', which is computed by equation (4). $s_{ij,i'j'}$ represents the similarity between edge e_{ij} and $e_{i'j'}$:

$$s_{ij,i'j'} = \exp(-\frac{1}{2}(\frac{e_{ij} - e_{i'j'}}{\sigma_e})^2) \tag{8}$$

where, e_{ij} and $e_{i'j'}$ denotes the lateral distance between feature i and j, and feature i' and j', as shown in Fig. 3. σ_e can be learned offline. The optimization problem will be solved by general random re-weighted walk framework [39].

Step 4: Feature tracking

The process builds association between features in consecutive frames. Since the perceived features are static and keep local structure, we formulate the process as a multi-order graph matching problem similar to equation (7).

Step 5: Temporal smoothing

The process constructs the optimal consistent matching between perceived features in consecutive frames and map features. Matching correctness of current frame can be verified by previous matching results in sliding window. Further, if a mismatch occurred in current frame, it can be found and corrected based on previous matching and tracking. Temporal smoothing acquires corresponding perceived feature s_i of map feature x^l by weighting the matching $D_{1:T}$ and matching confidence $c_{t,i}$ over each frame in sliding window:

$$\begin{aligned} \hat{s}_i &= \arg\max_{s_i} p(s_i|D_{1:T}) \\ &= \arg\max_{s_i} \frac{\sum_t I(s_i, D_t) c_{t,i}}{\sum_{t,j} I(s_j, D_t) c_{t,j}} \end{aligned} \tag{9}$$

where, $I(s_i, D_t)$ denotes if map feature x^l is matched with perceived feature s_i. The matching confidence $c_{t,i}$ is given by evaluating feature and local structure similarity:

$$\begin{aligned} c_{t,i} =& \omega \exp(-\frac{1}{2}(\frac{s_{ii'}}{\sigma_p})^2) \\ &+(1-\omega)\exp(-\frac{1}{2}(\frac{\frac{1}{N_m-1}\sum_{j=1,j\neq i}^{N_m} s_{ij,i'j'}}{\sigma_e})^2) \end{aligned} \tag{10}$$

If the accumulated confidence of the best perceived feature is much larger than that of the second best perceived feature, the best perceived feature will be considered as the matching pair of map feature x^l. Otherwise, the map feature x^l is considered to have uncertain matching, and the matching probability with each perceived feature can be given. The process distinguishes certain and uncertain matching, which can solve the problem of mismatching caused by singularity.

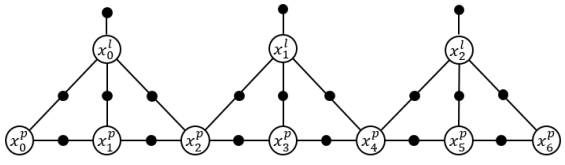
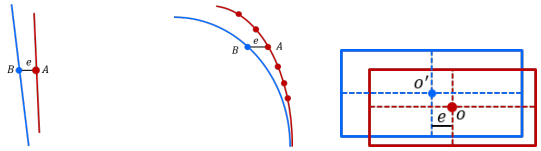


Fig. 4. Factor graph representation of the localization problem.



(a) Pole (b) Roadline (c) Traffic Sign

Fig. 5. Illustration of reprojection error model. Perceived features are denoted in red, reprojection features are expressed in blue.

C. Pose Graph Optimization

The pose estimation process of equation (2) can be defined as the product of priori probability and likelihood:

$$\begin{aligned} \hat{\mathcal{X}} &= \arg\max_{\mathcal{X}} p(\mathcal{X}|\mathcal{Z}, \mathcal{L}, \hat{\mathcal{D}}) \\ &= \arg\max_{\mathcal{X}} p(\mathcal{X}) p(\mathcal{Z}|\mathcal{X}, \mathcal{L}, \hat{\mathcal{D}}) \end{aligned} \tag{11}$$

By Gaussian distribution hypothesis, the priori distribution $p(\mathcal{X})$ is obtained through relative motion estimation of odometry. We formulate a sliding window non-linear least square estimator based on odometry measurement $z^o_{i,i+1}$ and matching pair of feature z^l_i to estimate the recent T poses. In comparison to commonly used filter methods, optimization method can deal with asynchronous and delayed measurements, and achieve higher accuracy with the same computing resources [40]. The optimization objective is represented as:

$$\begin{aligned} \hat{\mathcal{X}} = \arg\min_{\mathcal{X}} & \sum_i e^o(x^p_i, x^p_{i+1}, z^o_{i,i+1})^{\mathrm{T}} \Omega^o_i \\ & e^o(x^p_i, x^p_{i+1}, z^o_{i,i+1}) \\ & + \sum_i e^l(x^p_i, x^l, z^l_i)^{\mathrm{T}} \Omega^l_i e^l(x^p_i, x^l, z^l_i) \\ & + \sum_i e^m_j(x^l)^{\mathrm{T}} \Omega^m_j e^m_j(x^l)) \end{aligned} \tag{12}$$

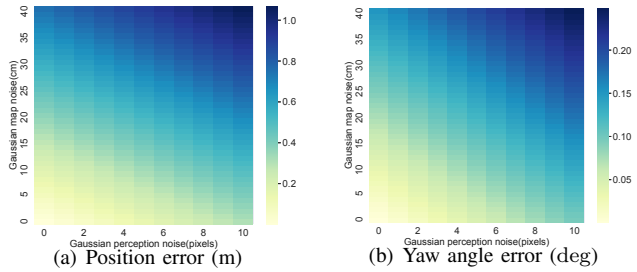
where, each error term together with corresponding information matrix can be regarded as a factor, each state variable can be regarded as a node, therefore the localization problem can be expressed by factor graph, as shown in Fig. 4. Error terms are composed of odometry error e^o, semantic measurement error e^l and map error e^m_j. The odometry error is defined as:

$$e^o(x^p_i, x^p_{i+1}, z^o_{i,i+1}) = ({x^p_{i+1}}^{\mathrm{T}} x^p_i) z^o_{i,i+1} \tag{13}$$

Semantic measurement error factor e^l is expressed as:

$$e^l(x^p_i, x^l, z^l_i) = [\frac{1}{Z^c_i} K T^c_v {x^p_i}^{\mathrm{T}} x^l]_0 - [z^l_i]_0 \tag{14}$$

where, $[\cdot]_0$ denotes the first element of a vector. The measurement error adopts only lateral error to eliminate the influence



(a) Position error (m) (b) Yaw angle error (deg)

Fig. 6. Localization performance under different perception and map errors.

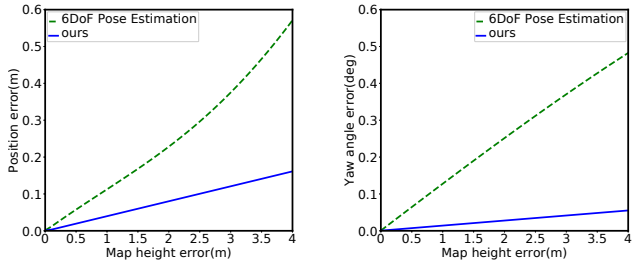


Fig. 7. Localization performance under different map height errors.

of height error and the requirements of accurate absolute height for map features, as shown in Fig. 5.

Map error factor e_j^m is represented as:

$$e_j^m(x^l) = x^l - m_j \tag{15}$$

where, m_j is the position of jth map feature. In this paper, we adopt the variance construction method of map features proposed in reference [26]. With an isotropic assumption of map factors, according to the assumed map quality, the variance of map factors can be defined as:

$$\sigma_m^2 = \frac{1}{\gamma(c)} r^2 \tag{16}$$

where, γ is an inverse-chi-squared cumulative distribution function, c denotes to the confidence, r represents the radius.

The non-linear optimization problem can be solved directly by an iterative algorithm. Sliding window instead of full batch method is adopted to improve computing efficiency while ensuring positioning accuracy. Old states are truncated and ignored directly. Marginalization method can also deal with old states, but it accumulates linearization errors, makes system matrix dense, and causes deadlock. Marginalization method constraints pose based on past data, but using map features as prior sufficiently constrains vehicle pose.

V. EXPERIMENTAL EVALUATION

A. Simulation in typical scenes

We first evaluate how perception and map errors affect the performance of the proposed algorithm. A typical intersection scene with four poles placed on both sides of the road, two traffic lights and two traffic signs in front of the vehicle is selected. The position and yaw angle errors under different perception and map errors are illustrated in Fig. 6. All results are the average of 1000 trials to eliminate randomness. It

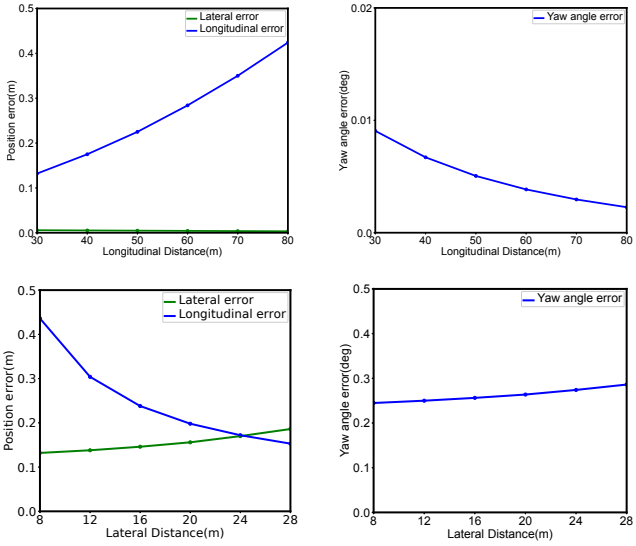


Fig. 8. Localization performance in various scenes with different semantic feature distributions.

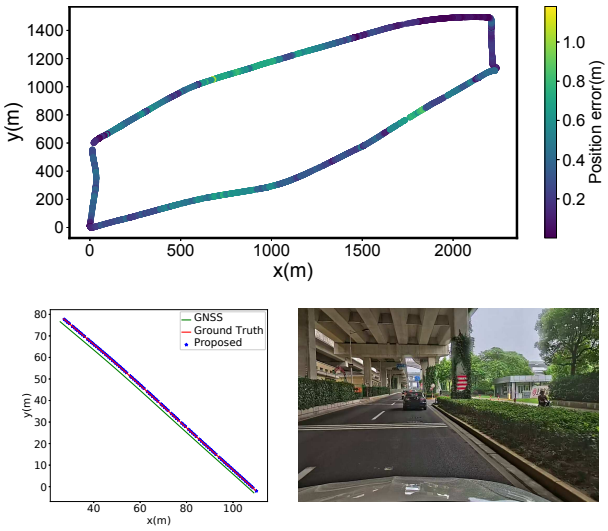


Fig. 9. Trajectory with localization errors of our 30km dataset and the comparing result of our approach against GNSS in elevated road scene.

shows that the proposed approach obtains highly accurate localization with 0.23m position error and 0.064° yaw error under 0.2m map error and 5 pixels perception error.

The proposed algorithm does not require highly accurate absolute height for map features. We compare our method with 6DoF pose estimation method to reveal the influence of different height errors of map features. The experiment is also implemented on the typical intersection. Fig. 7 illustrates the results, it demonstrates that our method obtains significantly smaller error for both position and yaw angle. Therefore our lateral error measurement model is beneficial for the performance of proposed system.

Different feature distributions can also affect localization performance. To explore the performance limitations of the proposed algorithm, we quantify it in various scenes with different feature distributions. We divide them into two cases,

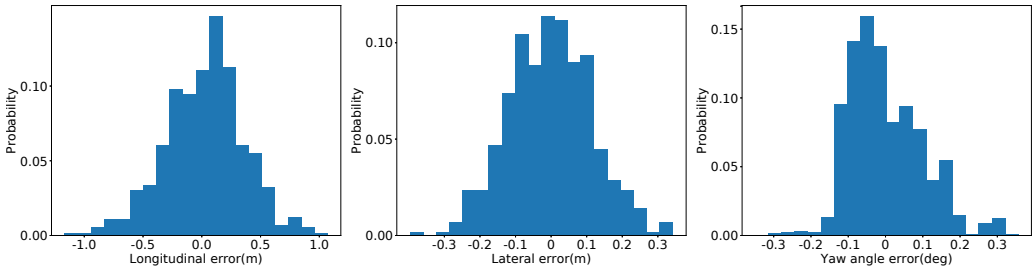


Fig. 10. Localization error distributions of the proposed algorithm.

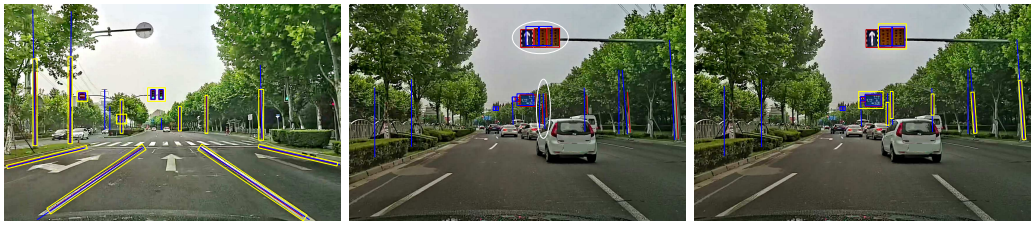


Fig. 11. Illustration of data association results. Perceived features are denoted in red, reprojection features are expressed in blue. Yellow boxes indicate successfully matched features. White ellipses represent ambiguity of matching, it is resolved by the proposed DA method.

including different longitudinal distances between vehicle and features, and different lateral distances between features in vehicle coordinate system. A typical scene with two poles on both sides of the vehicle with different distances is selected. Map and perceived features are assumed with a standard deviation of 0.05m and 2 pixels respectively. The experimental results are presented in Fig. 8. It illustrates that the longer the longitudinal distance, the worse the longitudinal positioning and the better lateral and yaw positioning. Whereas, the longitudinal accuracy gets better but the lateral and yaw positioning get worse as lateral distance increases.

B. Real Uran Road Scene

We evaluate the performance of the proposed algorithm with our vehicle on a 30km urban road with various scenes such as urban canyons, tunnels and viaducts. To evaluate positioning accuracy, the localization result fusing LiDAR, RTK and odometry is treated as the groundtruth. The localization performance of the proposed system compared with a Nearest Neighbor (NN) based 6DoF pose estimation method in various scenes is shown in Table I, it depicts that the proposed system achieves higher accuracy in various scenes with a mean longitudinal error e_{lo} of 0.43m, a mean lateral error e_{la} of 0.12m and a mean yaw angle error e_y of 0.11°. Fig. 9 gives the trajectory with localization errors and compares the localization accuracy with GNSS on elevated roads. It demonstrates that our approach has significantly lower error, which proves the effectiveness of our approach in complex road scene which GNSS fails. Fig. 10 gives the lateral, longitudinal and yaw error distributions.

The correctness of the proposed DA method is also evaluated with experiments in various scenes with false and missed detections. Fig. 11 gives the DA results of two scenes. The left image shows the correct matching of our approach in general scenes. The middle and right images illustrate that the association method deals with ambiguity caused by false detections effectively, this mainly benefits from the local structural and temporal consistency constraints.

TABLE I

LOCALIZATION PERFORMANCE OF PROPOSED ALGORITHM.

Scenes	Mean Localization Error					
	Ours			NN based 6DoF method		
	e_{lo}(m)	e_{la}(m)	e_y(deg)	e_{lo}(m)	e_{la}(m)	e_y(deg)
All	0.43	0.12	0.11	0.62	0.15	0.16
General	0.41	0.10	0.11	0.60	0.14	0.15
Urban canyons	0.46	0.13	0.11	0.65	0.16	0.16
Viaducts	0.45	0.13	0.11	0.64	0.16	0.17

We also compare the size of HD map against feature map and point cloud map. HD map only requires approximately 10KB per kilometer while 53MB and 3MB of feature map and compressed point cloud map.

VI. CONCLUSIONS

In this paper we presented a novel semantic localization algorithm that exploits semantic features extracted from image and features in HD map. To deal with the ambiguity of DA caused by repeated structures, missed and false detections, a robust DA method considering local structural, global pattern and temporal consistency was performed. A sliding window factor graph optimization framework that tightly couples association and odometry measurements was designed for accurate and robust localization. The proposed system was validated, and the results illustrate that the proposed approach achieves submeter-level localization accuracy.

In the future, we want to extend our framework with a more comprehensive theory to assess localization performance in real-time. We also plan to detect more semantic features to improve the robustness of the proposed system.

REFERENCES

[1] B. Eissfeller, G. Ameres, V. Kropp, and D. Sanroma, "Performance of gps, glonass and galileo," in *Photogrammetric Week*, vol. 7, 2007, pp. 185–199.

[2] P. Misra and P. Enge, "Global positioning system: signals, measurements and performance second edition," *Global Positioning System: Signals, Measurements And Performance Second Editions*, vol. 206, 2006.

[3] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[4] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[5] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual-inertial and multi-map slam," *arXiv preprint arXiv:2007.11898*, 2020.

[6] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[7] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[8] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, 2007, pp. 3565–3572.

[9] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and Systems*, vol. 2, no. 9, 2014.

[10] T. Shan and B. Englot, "Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4758–4765.

[11] B. M. Scherzinger, "Precise robust positioning with inertial/gps rtk," in *Proceedings of the 13th International Technical Meeting fo the Satellite Division of the Institute of Navigation (ION GPS)*. Citeseer, 2000, pp. 115–162.

[12] F. Caron, E. Duflos, D. Pomorski, and P. Vanheeghe, "Gps/imu data fusion using multisensor kalman filtering: introduction of contextual aspects," *Information fusion*, vol. 7, no. 2, pp. 221–230, 2006.

[13] M. Schreiber, H. Königshof, A.-M. Hellmund, and C. Stiller, "Vehicle localization with tightly coupled gnss and visual odometry," in *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 858–863.

[14] K. Jo, K. Chu, and M. Sunwoo, "Interacting multiple model filter-based sensor fusion of gps with in-vehicle sensors for real-time vehicle positioning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 1, pp. 329–343, 2011.

[15] R. P. D. Vivacqua, M. Bertozzi, P. Cerri, F. N. Martins, and R. F. Vassallo, "Self-localization based on visual lane marking maps: An accurate low-cost approach for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 2, pp. 582–597, 2017.

[16] M. Sons, M. Lauer, C. G. Keller, and C. Stiller, "Mapping and localization using surround view," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 1158–1163.

[17] P. Mühlfellner, M. Bürki, M. Bosse, W. Derendarz, R. Philippsen, and P. Furgale, "Summary maps for lifelong visual localization," *Journal of Field Robotics*, vol. 33, no. 5, pp. 561–590, 2016.

[18] M. Bürki, M. Dymczyk, I. Gilitschenski, C. Cadena, R. Siegwart, and J. Nieto, "Map management for efficient long-term visual localization in outdoor environments," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 682–688.

[19] M. Bürki, I. Gilitschenski, E. Stumm, R. Siegwart, and J. Nieto, "Appearance-based landmark selection for efficient long-term visual localization," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4137–4143.

[20] M. Bürki, C. Cadena, I. Gilitschenski, R. Siegwart, and J. Nieto, "Appearance-based landmark selection for visual localization," *Journal of Field Robotics*, vol. 36, no. 6, pp. 1041–1073, 2019.

[21] M. Bürki, L. Schaupp, M. Dymczyk, R. Dubé, C. Cadena, R. Siegwart, and J. Nieto, "Vizard: Reliable visual localization for autonomous vehicles in urban outdoor environments," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1124–1130.

[22] M. Schreiber, C. Knöppel, and U. Franke, "Laneloc: Lane marking based localization using highly accurate maps," in *2013 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2013, pp. 449–454.

[23] F. Poggenhans, N. O. Salscheider, and C. Stiller, "Precise localization in high-definition road maps for urban regions," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2167–2174.

[24] Y. Lu, J. Huang, Y.-T. Chen, and B. Heisele, "Monocular localization in urban environments using road markings," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 468–474.

[25] J. Jeong, Y. Cho, and A. Kim, "Road-slam: Road marking based slam with lane-level accuracy," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 1736–1473.

[26] D. Wilbers, C. Merfels, and C. Stachniss, "Localization with sliding window factor graphs on third-party maps for automated driving," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5951–5957.

[27] R. Spangenberg, D. Goehring, and R. Rojas, "Pole-based localization for autonomous vehicles in urban scenarios," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2161–2166.

[28] M. Sefati, M. Daum, B. Sondermann, K. D. Kreisköther, and A. Kampker, "Improving vehicle localization using semantic and pole-like landmarks," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 13–19.

[29] C. Wu, T. A. Huang, M. Muffert, T. Schwarz, and J. Gräter, "Precise pose graph localization with sparse point and lane features," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 4077–4082.

[30] W.-C. Ma, I. Tartavull, I. A. Bârsan, S. Wang, M. Bai, G. Mattyus, N. Homayounfar, S. K. Lakshmikanth, A. Pokrovsky, and R. Urtasun, "Exploiting sparse semantic hd maps for self-driving vehicle localization," *arXiv preprint arXiv:1908.03274*, 2019.

[31] J. Kümmerle, M. Sons, F. Poggenhans, T. Kühner, M. Lauer, and C. Stiller, "Accurate and efficient self-localization on roads using basic geometric primitives," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5965–5971.

[32] J. Jin, X. Zhu, Y. Jiang, and Z. Du, "Localization based on semantic map and visual inertial odometry," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2410–2415.

[33] Z. Xiao, D. Yang, T. Wen, K. Jiang, and R. Yan, "Monocular localization with vector hd map (mlvhm): A low-cost method for commercial ivs," *Sensors*, vol. 20, no. 7, p. 1870, 2020.

[34] H. Hu, M. Sons, and C. Stiller, "Accurate global trajectory alignment using poles and road markings," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1186–1191.

[35] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[36] Y. Bar-Shalom and E. Tse, "Tracking in a cluttered environment with probabilistic data association," *Automatica*, vol. 11, no. 5, pp. 451–460, 1975.

[37] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1722–1729.

[38] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[39] J. Lee, M. Cho, and K. M. Lee, "Hyper-graph matching via reweighted random walks," in *CVPR 2011*. IEEE, 2011, pp. 1633–1640.

[40] D. Wilbers, C. Merfels, and C. Stachniss, "A comparison of particle filter and graph-based optimization for localization with landmarks in automated vehicles," in *2019 Third IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2019, pp. 220–225.