Landmark based localization in urban environment

Article *in* ISPRS Journal of Photogrammetry and Remote Sensing · September 2017

DOI: 10.1016/j.isprsjprs.2017.09.010

CITATIONS
15

READS
831

3 authors:

Xiaozhi Qu
Didi chuxing Technology
9 PUBLICATIONS **105** CITATIONS
SEE PROFILE

Bahman Soheilian
University of Paris-Est
44 PUBLICATIONS **729** CITATIONS
SEE PROFILE

Nicolas Paparoditis
Institut national de l'information géographique et forestière
172 PUBLICATIONS **2,160** CITATIONS
SEE PROFILE

Landmark based localization in urban environment

Xiaozhi Qu*, Bahman Soheilian, Nicolas Paparoditis

Univ. Paris-Est, LASTIG MATIS, IGN, ENSG, F-94160 Saint-Mande, France

Abstract

A landmark based localization with uncertainty analysis based on cameras and geo-referenced landmarks is presented in this paper. The system is developed to adapt different camera configurations for six degree-of-freedom pose estimation. Local bundle adjustment is applied for optimization and the geo-referenced landmarks are integrated to reduce the drift. In particular, the uncertainty analysis is taken into account. On the one hand, we estimate the uncertainties of poses to predict the precision of localization. On the other hand, uncertainty propagation is considered for matching, tracking and landmark registering. The proposed method is evaluated on both KITTI benchmark and the data acquired by a mobile mapping system. In our experiments, decimeter level accuracy can be reached.

Keywords: Visual odometry; landmark; multi-camera; local bundle adjustment; uncertainty

1. Introduction

Precise localization is desired for many applications such as MMS (Mobile Mapping System), self-driving cars, ADAS (Advanced Driving Assistance Systems) and AR (Augmented Reality). The solutions for localization can be summarized as global localization and position tracking (Thrun et al., 2005). Global localization measures absolute positions, while position tracking starts from a point and tracks the relative poses over time. The most famous global localization system is GNSS (Global Navigation Satellite System). However, the multi-path or mask of satellite signals could lead to large errors and even outage of localization in the urban environments. Moreover,

*Corresponding author

GNSS provides positioning once a second and the orientations can not be measured directly. Therefore, some direct positioning and orientation systems are developed by combining GNSS with INS (Inertial Navigation System). The gaps between GNSS points are bridged with INS measurements, which improves positioning rate and compensates the errors caused by GNSS multi-path or mask. This kind of solutions are mature enough and widely used on MMS and autonomous navigation. However, drift is an innate issue for INS and high-quality INSs confine the drift in return of high cost.

A more affordable solution is SLAM (Simultaneous Localization and Mapping), which is able to use low-cost sensors. In practice, cameras are the most widely used sensor for SLAM. Compared with some active sensors like laser, cameras are cheaper and rich information (e.g. texture, spatial) can be acquired. The solutions for SLAM using cameras are usually called Visual Odometry (VO), which estimate relative poses by tracking the correspondences across images (Nistér et al., 2004). Different techniques were proposed to improve the accuracy of VO and they can be classified as the probabilistic filter (e.g. Extended Kalman Filter (EKF), Particle Filter (PF)) and Bundle Adjustment (BA). BA achieves better accuracy than EKF or PF (Ji and Yuan, 2016), because BA provides a global least squares optimization and adjusts image poses and object points by taking into account the entire observation equations and eventual constraints. However, the size of equation system increases rapidly with the growing number of images. Then, Local Bundle Adjustment (LBA) is proposed to reduce the complexity of BA (Mouragnon et al., 2006). It employs BA on a fixed size of sliding window and propagates the uncertainties through image sequence (Eudes and Lhuillier, 2009). SLAM is a solution for position tracking, the localization is conducted on a local system and errors accumulate over time. Although the global drift can be reduced by loop closure, there is no loop in many scenarios for a moving robot or vehicle in urban areas and computation of large loops is time-consuming.

Many methods were proposed to integrate GNSS with VO, where the drift was compensated using the positions measured by GNSS (Agrawal and Konolige, 2006; Wei et al., 2011; Lhuillier, 2012). However, the accuracy depends on the precision of GNSS data and GNSS measurements are not always reliable in urban environment. The drift increases quickly when the system suffers multi-path or mask problems. Therefore, more external data needs to be considered for precise localization based on cameras.

2. Related work

Nowadays, many VO methods have been proposed and different types of maps have been designed, produced and applied for precise localization. In this section, both state-of-the-art VO approaches and map based localization methods are investigated, our landmark based localization approach is introduced briefly at the end.

2.1. Visual odometry

The visual odometry was firstly proposed by Nister (Nistér et al., 2004). Current VO approaches can be summarized as feature based and direct VO approaches. Most feature based VO methods follow the approach introduced in PTAM (Parallel Tracking and Mapping) (Klein and Murray, 2009), which has three main modules: (1) feature extraction, matching and tracking; (2) poses and map points estimation and optimization; (3) loop closure and global optimization. Usually, interest points or lines are detected in images and matched over sequences. The optimal solutions are achieved by minimizing back projection errors with the methods such as bundle adjustment, EKF and PF. PTAM system was designed for indoor environment and the computational cost increases quickly in large scale environment. A more efficient and accurate approach on feature based VO, is proposed as ORB-SLAM (Mur-Artal and Tardós, 2016). It detects the ORB (Oriented FAST and Rotated BRIEF) features in images for matching and tracking. The local map is optimized using local bundle adjustment. The loop is recognized based on DBoW2 (Galvez-Lpez and Tardos, 2012) built on ORB features, and all the poses are optimized using global bundle adjustment once the loop is detected successfully.

Different with feature based VO methods which aim at minimizing back-projection errors, direct approaches usually minimize the photometric errors on image intensities (Forster et al., 2014). A representative work on direct VO was LSD-SLAM (Engel et al., 2014), it built semi-dense maps in large scale, based on real-time image alignment. A recent work of Engel et al (Engel et al., 2016) known as DSO (Direct Sparse Odometry), had the benefits from both direct (no feature extraction) and sparse (joint optimization of parameters), so DSO achieved better accuracy than LSD-SLAM. Nevertheless, theses direct VO methods still need features for loop closure (Mur-Artal and Tardós, 2016). Loop closure can compensate the drift of VO over time, but the absolute scale is very difficult to determine using mono based VO.

Moreover, the paths of a moving agent do not always have loops in urban environment. In this case, geo-referenced maps need to be integrated into VO system for precise localization.

2.2. Map based localization methods

In this paper, we classify the map based localization methods according to the types of the map, which are *low level maps*, *conventional maps* and *semantic maps*.

2.2.1. Low level maps

The low level maps are composed of geo-referenced visual features or point clouds. They are usually used for localization in a *teach and repeat* (e.g. route following for robot) (Furgale and Barfoot, 2010). The maps are built in teaching steps with SFM (Structure From Motion) (Royer et al., 2007) or scanning point clouds by a MMS (Bodensteiner et al., 2011). In a repeating step, the robot intends to follow the same route by registering the images captured by on-board cameras with maps, according to the similarity of visual features or intensity maps for point clouds.

Low level maps are easy to produce, but they have two main drawbacks. First, the high storage volume is required due to rich features extracted from images or point clouds acquired by a laser scanner. Second, incremental updates are extremely complicated. For instance, it is difficult to know which 3D points are changed.

2.2.2. Conventional maps

Conventional maps represent generalized models of the world and they are composed of elements like segments, polygons and planes. Light storage is needed and they are easy to be updated in comparison to low level maps. Thus, more methods integrate conventional maps (e.g. OpenStreetMap, DEM (Digital Elevation Model), 3D models) with VO. For 2D maps like OpenStreetMap, curve-to-curve map matching can be applied to correct the drift for localization on road, because the road segments, building boundary, locations and attributes of objects are known. These methods usually use a cheap GPS for initial positions (Alonso et al., 2012; Brubaker et al., 2013). Recently, a graph matching based method was proposed to align the trajectory estimated by VO with 2D road network without initial positions (Gupta et al., 2016). For higher dimension conventional maps such as DEM and 3D city model, direct pose estimation is possible. For instance, the pose of an

individual image is optimized by registering images with 2.5D untextured city model based on semantic image segmentation (Arth et al., 2015). A more direct strategy is to register the coarse structure generated by VO with 3D city model to correct the drift (Lothe et al., 2009) and even to estimate poses by registering images with textured 3D model based on mutual information (Caron et al., 2014). However, the precision of conventional maps is 1-5 meters, which is not sufficient for precise localization.

2.2.3. Semantic maps

In urban areas, there are rich semantic features like road marks, road signs and other man-made objects which are well-defined objects. They can be detected and reconstructed precisely from ground based imagery (Soheilian et al., 2010), aerial images (Tournaire et al., 2006) and point clouds (Hervieu et al., 2015). In particular, these semantic objects are easy to be detected in images that make them be convenient to integrate in VO approaches. Different types of objects were integrated for precise localization. The lane markings extracted from aerial images, were used to correct positioning errors by aligning the lane marks detected in images with maps based on ICP (Iterative Closest Point) algorithm (Pink, 2008). A similar strategy was proposed, but the precise locations were estimated via map matching between pre-built road mark maps and road marks extracted from stereo images (Schreiber et al., 2013). These two methods need initial positions measured by a GPS. Moreover, the vehicle poses can also be optimized with the constraints generated by 2D-3D correspondences between images and geo-referenced road signs (Wei et al., 2014) or road marks (Tournaire et al., 2006). These semantic objects can be expressed using segments or polygons in the database, so they are convenient to be stored and updated.

2.3. Our strategy

A featured based VO system is developed in this paper. In our previous work, we integrated geo-referenced road signs (Qu et al., 2015) and road marks (Soheilian et al., 2016) with VO using a single camera. The road signs were reconstructed from images (Soheilian et al., 2013) and the road marks were extracted from point clouds (Hervieu et al., 2015) captured by a MMS. The localization was optimized with LBA and some Ground Control Points (GCPs) were generated by registering road signs or road marks in a pre-built database in order to reduce the drift. Besides, uncertainties were estimated to predict the uncertainty of poses. In this paper, we regard geo-referenced

road signs and road marks as landmarks and they are integrated for multiple camera configurations. Fig. 1 shows the working flow of the proposed

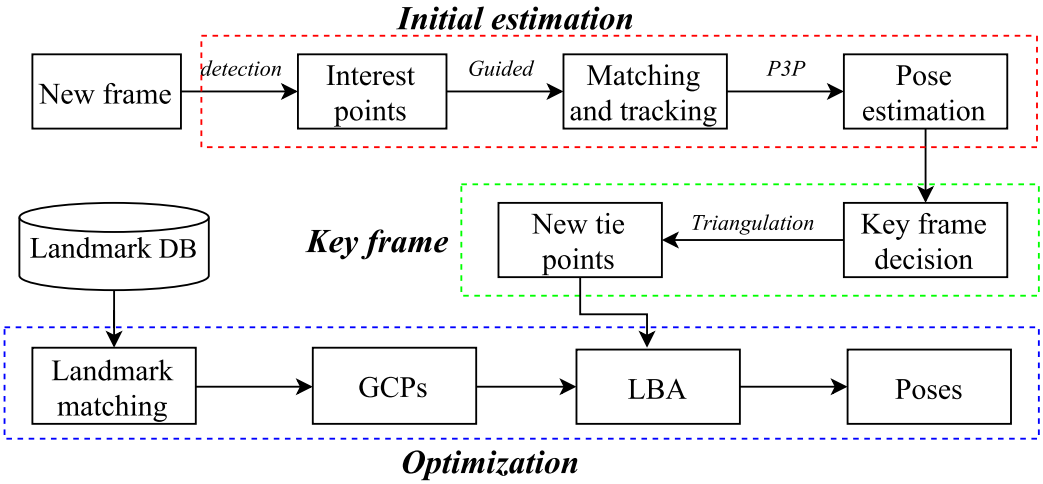


Figure 1: Flowchart of landmark based localization.

method. There are three main parts in the system: initial pose estimation for each frame, key frame selection and integration of landmarks for optimization. The proposed system starts from a known point which can be provided using low-cost GPS. Even though the initial point is not very accurate, the system can still handle the uncertain initialization given its uncertainty. Indeed this uncertainty can be propagated to the trajectory and be resolved after integrating the geo-referenced landmarks. For each new image, interest points are detected, matched and tracked over the sequence, this will be introduced in section 3. The pose is estimated for every frame, but only key frames are selected for LBA. Section 4 will present the pose estimation and optimization. The landmarks are registered to generate GCPs for LBA. This part will be introduced in section 5. Finally, the proposed method will be validated using real images, where multiple camera configurations and the improvement provided by landmark integration will be tested.

There are three new contributions in comparison to our previous work. First, a new matching and tracking method is proposed to improve both accuracy and efficiency. Second, the localization is extended to adopt any camera configuration such as monocular, stereo, non-overlap stereo and multi-camera. Finally, our method is evaluated on KITTI benchmark and real image sequences captured by a MMS developed at IGN (National Institute

of Geographic Information and Forestry).

3. Propagation based matching and tracking

Many SFM approaches detect interest points in an image and then match each point in the reference image by searching target image entirely, but this kind of strategies are not efficient enough for VO. Besides, many false matches will be produced in repeatable texture areas such as road and building facade. In VO or SLAM approaches, the searching of visual features are associated with a *prediction* procedure. Davison formulated the problem as an EKF (Davison, 2003). Searching region was determined according to the uncertainty propagation of 3D map points and predicted poses. However, it is difficult to maintain the covariances of maps for a large scene because of the quick growth of the matrix size. In ORB-SLAM, a more efficient strategy is proposed for feature tracking, where the pose of the new frame is predicted using a constant velocity model according to the states of prior poses, thus the searching of correspondences is limited in a small region in terms of the projection of 3D map points in image (Mur-Artal and Tardós, 2016). The searching size is determined according to the location of each map point relative to the image and is scaled with the predicted scale factor of the map point from the previous image because the image points are detected in image pyramid space for ORB key points. We adopted a similar procedure as ORB-SLAM, but there are two differences in our method: (1) interest points are detected with the FAST detector and we do not need descriptors. (2) the searching area is determined according to uncertainty propagation of 3D map point and pose prediction.

Fig. 2 shows the work-flow of proposed matching and tracking method. There are three main steps that will be introduced in following three sections.

3.1. Pose prediction and uncertainty propagation

A velocity-constant model is used for prediction. The pose at time step t is noted as P_t which is a 6×1 vector (position and three Euler angles), thus the pose at time step $t+1$ is computed by:

$$P_{t+1}^{*} = P_t + V_t \cdot \Delta t \tag{1}$$

where Δt is time interval from t to $t+1$ and P_{t+1}^{*} is the predicted pose at time step $t+1$. V_t stands for velocity vector and it is defined as the average

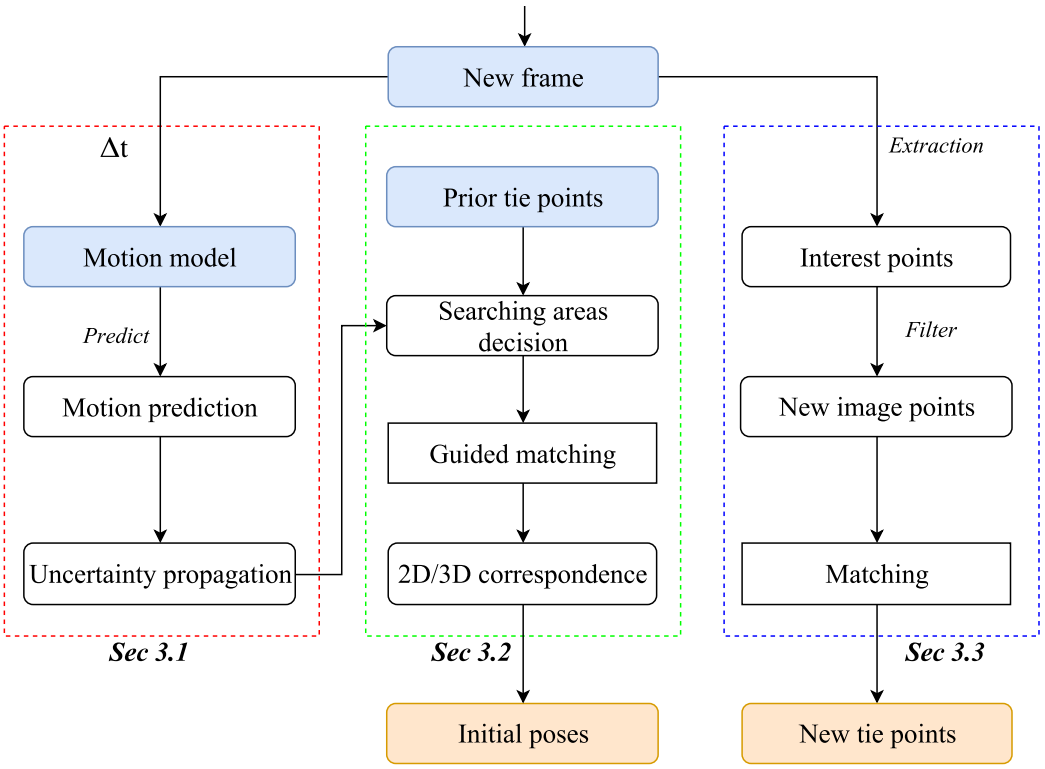


Figure 2: Flowchart of propagation based matching and tracking.

velocity from $t-2$ to t:

$$V_t = \frac{(P_t - P_{t-1}) + (P_t - P_{t-2})}{2\Delta t_1 + \Delta t_2} \tag{2}$$

in which $\Delta t_1 = T_t - T_{t-1}$ and $\Delta t_2 = T_{t-1} - T_{t-2}$, where, T represents the time. Let's define $\Delta T = 2\Delta t_1 + \Delta t_2$, then V_t can be written as:

$$V_t = \underbrace{\begin{bmatrix} \frac{2}{\Delta T}\boldsymbol{I}_{6\times6} & -\frac{1}{\Delta T}\boldsymbol{I}_{6\times6} & -\frac{1}{\Delta T}\boldsymbol{I}_{6\times6} \end{bmatrix}}_{\boldsymbol{A}} \begin{bmatrix} P_t \\ P_{t-1} \\ P_{t-2} \end{bmatrix}, \tag{3}$$

in which $\boldsymbol{I}_{6\times6}$ is a 6×6 identity matrix. At the beginning of localization, we suppose that V_t is zero, thus the pose of second frame is $P_1^* = P_0$. This assumption makes sense when the state change of vehicle is small. In practice, we can use an odometer for initialization.

Considering covariance propagation principle, the covariance matrix of V_t, noted as $\mathbf{\Sigma}_{V_t}$, can be estimated in term of Eq. 3:

$$\mathbf{\Sigma}_{V_t} = \boldsymbol{A}\mathbf{\Sigma}_P\boldsymbol{A}^T \tag{4}$$

where, $\mathbf{\Sigma}_P$ is the covariance matrix of P_t, P_{t-1} and P_{t-2}. Denote $\mathbf{\Sigma}^*_{P_{t+1}}$ as the covariance of P^*_{t+1}. Eq. 1 is a linear equation, then:

$$\mathbf{\Sigma}^*_{P_{t+1}} = \mathbf{\Sigma}_{P_t} + \Delta_t^2\mathbf{\Sigma}_{V_t} \tag{5}$$

3.2. Guided matching

With predicted pose P^*_{t+1}, the searching area of each tie point can be generated considering uncertainty propagation from P^*_{t+1} and tie points $\boldsymbol{X}$ (3D object points). Fig. 3 demonstrates our strategy. The searching area is

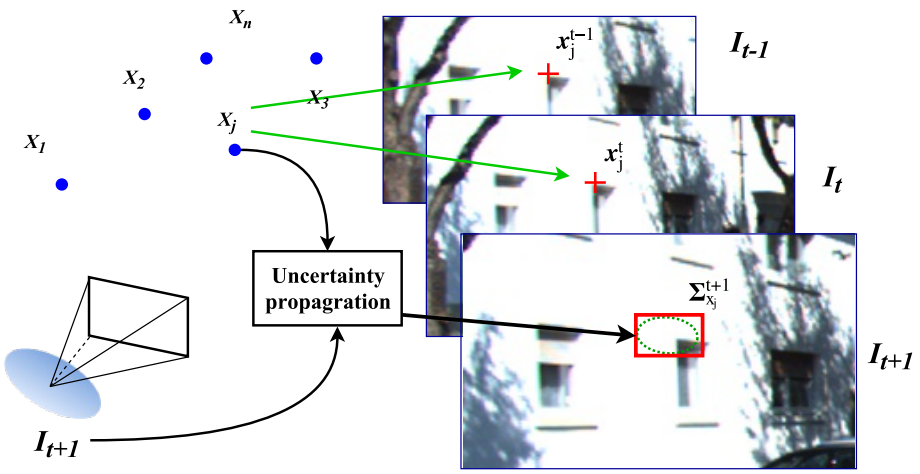


Figure 3: Generation of searching area. $X_1, ...X_j, ...X_n$ are tie points. The dotted ellipse in I_{t+1} is the error ellipse of the back projection of X_j and the red rectangle represents the searching area. I_t is reference image and I_{t+1} is target image.

the bounding box of error ellipse for the back-projection of tie point X_j in image. Let's define F as back-projection function for a perspective camera, hence 2D projection of X_j in image $t+1$ is :

$$x_j^{t+1} = F(P^*_{t+1}, X_j).$$

As F is a nonlinear function, thus the covariances of x_j^{t+1} is estimated from equation below:

$$\mathbf{\Sigma}_{x_j}^{t+1} = \begin{bmatrix} \frac{\partial F}{\partial_{P_{t+1}^*}} & \frac{\partial F}{\partial_{X_j}} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_{P_{t+1}^*} & 0 \\ 0 & \Sigma_{X_j} \end{bmatrix} \begin{bmatrix} \frac{\partial F}{\partial_{P_{t+1}^*}} \\ \frac{\partial F}{\partial_{X_j}} \end{bmatrix} \tag{6}$$

The precise location of tie point in new image is obtained by matching the reference image point in the searching area, that forms a template matching problem. The precise location of X_j in image $t+1$ is determined by maximizing the following Normalized Cross Correlation (NCC) score:

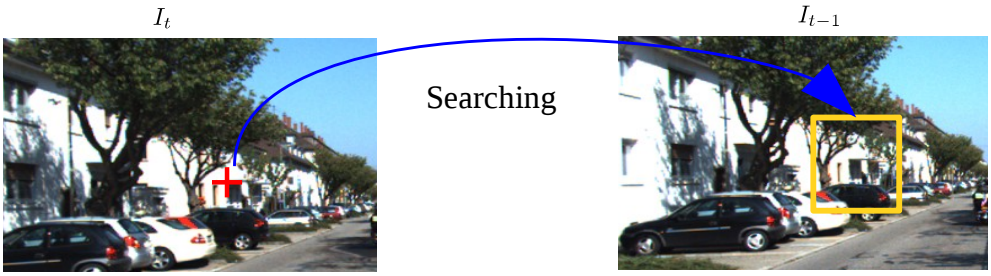
$$\hat{x}_j^{t+1} = \underset{x_j^{t+1} \in \Sigma_{x_j}^{t+1}}{\operatorname{argmax}} \{NCC(x_j^{t+1}, x_j^t)\} \tag{7}$$

Where, $x_j^{t+1} \in \Sigma_{x_j}^{t+1}$ stands for the range of x_j^{t+1} that is determined by the uncertain region of X_j in image $t+1$. Therefore, a set of 3D-2D correspondences can be obtained for estimating the pose of a new frame.

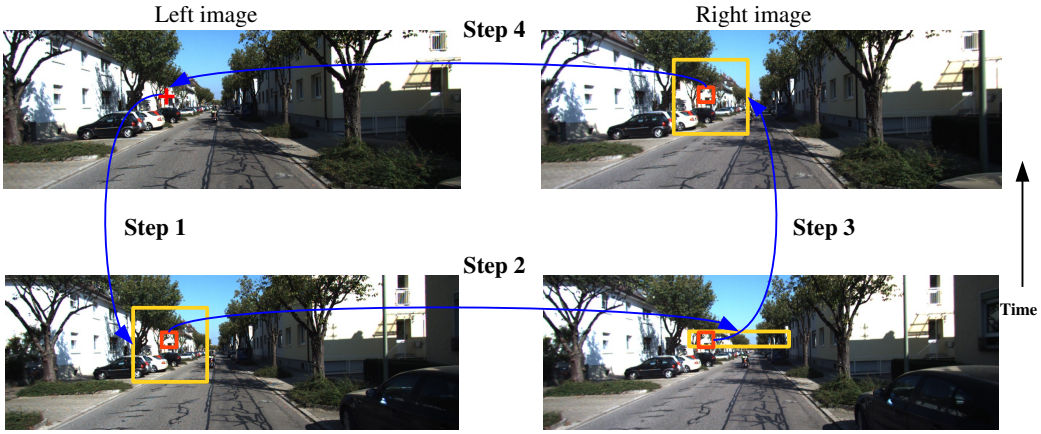
3.3. Searching new tie points

In order to maintain the state of the vehicle, new tie points should be reconstructed from new matches. To do this, new interest points are found in the new frame and then they are matched with the last frame to obtain the 2D correspondences. FAST (Features from Accelerated Segment Test) (Rosten and Drummond, 2006) is applied for interest points detection, but original FAST extracts too many points. We only keep the points when their minimal distances to existing interest points obtained by guided matching are larger than 20 pixels.

To match the new interest points with the last frame I_t, the correspondences between I_t and I_{t+1} obtained by guided matching are used as prior knowledge. Then, the bi-cubic interpolation method is used to predict the displacements of new image points in I_t. Fig. 4a shows the matching for monocular images, where the search area size is fixed as 20×20 pixels since no prior information about the uncertainty of scene is known and the center is determined according to the predicted displacements of a new interest point in I_t. A circle matching strategy is proposed for stereo sequences. There are four steps (*cf.* Fig. 4b). Step 1 and step 3 are the same as that proposed for monocular case, while step 2 is guided by the epipolar line. The final point is checked with the start point by computing NCC score at step 4. If the score is larger than 0.9, the circle matching result is accepted.



(a) Matching for monocular images



(b) Circle matching for consecutive stereo pairs

Figure 4: Matching for new image point. Red cross: new image point. Yellow square: fixed size of searching area.

4. Pose estimation and refinement

After matching and tracking, initial pose and new tie points can be estimated for the new frame. In this paper, some frames are selected out of the sequence according to some defined requirements for LBA and they are called *key frames*.

4.1. Initial estimation and key-frame selection

The estimation of initial pose of the new frame using the 3D-2D correspondences obtained by guided matching, is a PnP (Perspective n Point) problem (Quan and Lan, 1999). The minimal solution for PnP problem is P3P, which only needs three 3D-2D correspondences. A P3P algorithm proposed by Kneip et al. (2011) is applied in a RANSAC scheme for robust

pose estimation. For new tie points, initial 3D coordinates are computed by triangulation.

The initial pose for every frame is estimated, but only key frames are selected for LBA. Three conditions are defined for key frames selection:

- Tracking ratio is smaller than 0.3.
- Distance to the nearest key frame is longer than $1.5m$
- Rotating angle to the nearest key frame is larger than $10°$.

Here, the tracking ratio means the ratio between the number of images points obtained by guided matching and the total number of image points in a new frame. It varies between 0 and 1 and decreases with the displacement of camera. If a frame meets one of the conditions, it is a key frame.

4.2. Notations and definition for LBA

LBA process a sliding window including N frames. Last n frames ($n < N$) are newly estimated and the rest $N - n$ frames are inherited from previous steps. In LBA equations, we have following basic notations:

- $\boldsymbol{P}_n$: new poses in LBA window.
- $\boldsymbol{P}_p$: poses inherited from previous steps.
- $\boldsymbol{X}_t$: 3D tie point, expressed in world coordinate system.
- $\boldsymbol{v}_t$: back projection errors.
- $\boldsymbol{\Sigma}_t$: covariance matrix of image points.
- $\boldsymbol{v}_p$: residuals related to $\boldsymbol{P}_p$.
- $\boldsymbol{P}_p^0$: prior estimates of $\boldsymbol{P}_p$.
- $\boldsymbol{\Sigma}_p$: covariance matrix of $\boldsymbol{P}_p^0$, estimated by LBA in previous steps.

In LBA, $\boldsymbol{P}_p, \boldsymbol{P}_n, \boldsymbol{X}_t$ are parameters and they are expressed as:

$$\begin{array}{l} \boldsymbol{P}_p = [P_{t-N-1} \quad ... \quad P_{t-n}]^T \\ \boldsymbol{P}_n = [P_{t-n+1} \quad ... \quad P_t]^T \\ \boldsymbol{X}_t = [... \quad X_i \quad ...]^T \end{array} \tag{8}$$

Fig. 5 is an example to show the procedures of LBA, where $N = 5$, $n = 2$. From the second step, there are $N - n$ frames with priorly estimated poses and covariances since they have been resolved by LBA in previous steps. For instance, the poses of frames $2, 3, 4$ in Fig. 5 have been estimated at the first step, in the second step, these poses are regarded as parameters with prior knowledge. Whereas LBA at the first step is same as conventional BA.

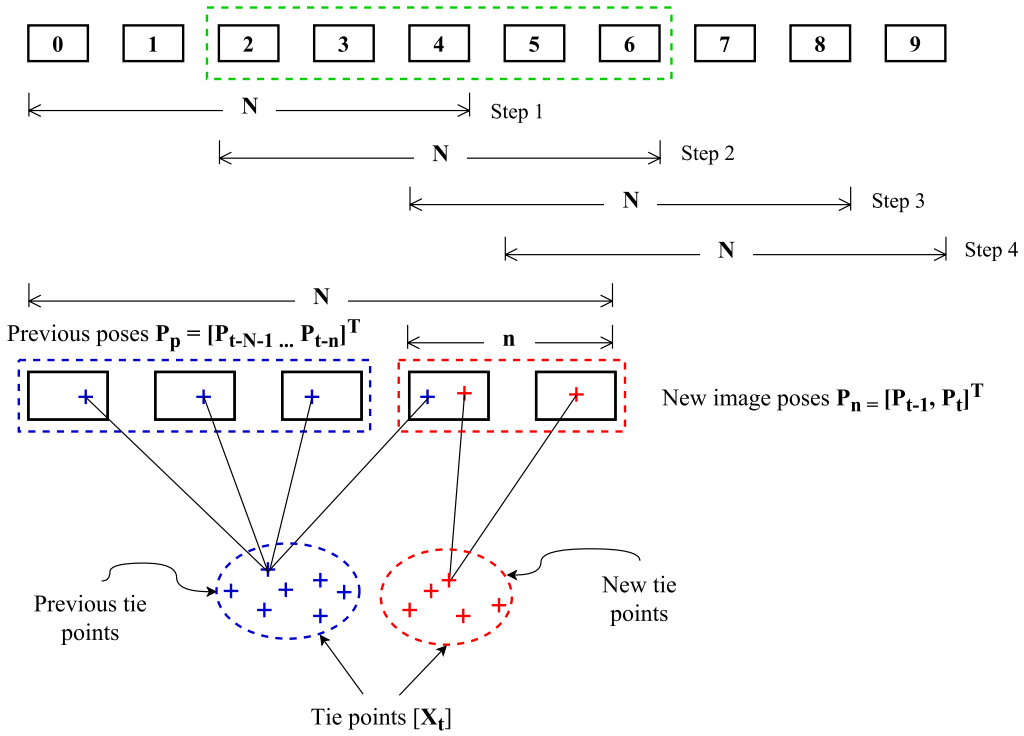


Figure 5: Schematic of the LBA procedure. The zoom-up diagram at left-bottom presents the details of second step marked with green dotted rectangle.

4.3. Mathematics of LBA

Conventional BA aims at minimizing the weighted sum of squared back-projection errors (Triggs et al., 2000). The back projection error is computed by:

$$\boldsymbol{v}_t = F(\boldsymbol{P}, \boldsymbol{X}_t) - \mathbf{x}_t \tag{9}$$

where $\mathbf{x}_t$ represents image points. Optimal poses and tie points can be obtained by minimizing the following cost function:

$$[\hat{\boldsymbol{P}}, \hat{\boldsymbol{X}}_t] = \underset{\boldsymbol{P}, \boldsymbol{X}_t}{\operatorname{argmin}} \{\frac{1}{2} \boldsymbol{v}_t^T \boldsymbol{\Sigma}_t^{-1} \boldsymbol{v}_t\} \tag{10}$$

in which, $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{X}}_t$ are optimal estimations.

Conventional BA achieves the most accurate estimates, but the computational complexity increases quickly with growing of frame number. In practice, VO or SLAM can be posed in terms of a graph problem. In order to reduce the computation, some prior poses can be marginalized out in a new processing window. The conventional marginalization can lead to fully inter-connected graph since every pose elimination causes new links between landmarks. The marginalization usually makes use of Schur-complement (Leutenegger et al., 2015), thus Hessian structure becomes dense which makes it harder to solve in bundle adjustment. In order to overcome this downside, the approaches proposed in DSO (Engel et al., 2016) or the visual-inertial odometry (Leutenegger et al., 2015) drop the residual terms which are not visible on the most recent key frame. This can keep the sparsity pattern of Hessian for efficient estimation. The LBA proposed in this paper keeps the key frame poses in a sliding window and the landmarks which can be observed by at least two images in the processing window. The landmarks which are not visible for the images in processing window, are dropped to reduce the computation. Uncertainties of prior poses are considered in LBA and a strong hypothesis is made that the prior poses are independent with landmarks. Compared with conventional BA, LBA can achieve acceptable covariance (Eudes and Lhuillier, 2009). The proposed LBA considers more connections between landmarks and poses, thus it can achieve better accuracy while the efficiency would be lower than the approaches using marginalization.

To separate the prior poses $\boldsymbol{P}_p$ and new poses $\boldsymbol{P}_n$ in F, we rewrite Eq. 9 as:

$$\boldsymbol{v}_t = F(\boldsymbol{P}_p, \boldsymbol{P}_n, \boldsymbol{X}_t) - \mathbf{x}_t \tag{11}$$

For LBA from second step, the uncertainties of prior poses will be considered. An extra error equation for $\boldsymbol{P}_p$ is added:

$$\boldsymbol{v}_p = \boldsymbol{P}_p - \boldsymbol{P}_p^0, \tag{12}$$

In this case, a new term is added into the cost function, the parameters are resolved by minimizing a joint cost function (Eudes and Lhuillier, 2009):

$$[\hat{\boldsymbol{P}}_p, \hat{\boldsymbol{P}}_n, \hat{\boldsymbol{X}}_t] = \underset{\boldsymbol{P}_p, \boldsymbol{P}_n, \boldsymbol{X}_t}{\operatorname{argmin}} \{\frac{1}{2}(\boldsymbol{v}_t^T \boldsymbol{\Sigma}_t^{-1} \boldsymbol{v}_t + \boldsymbol{v}_p^T \boldsymbol{\Sigma}_p^{-1} \boldsymbol{v}_p)\} \tag{13}$$

where, $\hat{\boldsymbol{P}}_p, \hat{\boldsymbol{P}}_n, \hat{\boldsymbol{X}}_t$ are optimal estimates. To solve the problem, an error array composed by Eq. 11 and 12 is linearized using first-order Taylor expansion. The optimal estimation for the parameters are achieved iteratively. In each iterative step, the parameter corrections are resolved from the normal equation:

$$\underbrace{(\boldsymbol{J}^T \boldsymbol{W} \boldsymbol{J})}_{\boldsymbol{H}} \hat{\boldsymbol{\delta}} = -\boldsymbol{J}^T \boldsymbol{W} \boldsymbol{y} \tag{14}$$

where, $\boldsymbol{H}$ is usually called Hessian matrix, $\boldsymbol{J}$ is the Jacobian matrix, $\hat{\boldsymbol{\delta}}$ stands for the corrections of parameters and $\boldsymbol{W} = diag(\boldsymbol{\Sigma}_p^{-1}, \boldsymbol{\Sigma}_t^{-1})$. The covariances of parameters can be obtained by $\boldsymbol{H}^{-1}$. The dimension of Hessian matrix is $6 \times N + 3 \times M$, where M is the number of tie points. In LBA, N is fixed, the computation of covariance matrix is dominated by M and $M \gg N$. In fact, the covariance matrix of image poses is the top-left block of $\boldsymbol{H}^{-1}$, thus, Schur complement is used to estimate the pose covariance matrix (Triggs et al., 2000).

4.4. LBA for multi-camera

In practice, VO always desires large FOV (Field Of View) and high angular resolution, because larger FOV enables to observe more informative areas and high angular resolution of the image has better angular accuracy on interest point detection, so that the accuracy of localization can be improved (Zhang et al., 2016). Using a single camera, a trade-off between angular resolution and FOV should be found. To overcome this, we combine multiple cameras together to enlarge FOV and retain high angular resolution. For multi-camera configuration, the matching and tracking of tie points is a combination of the methods introduced in section 3 for mono and stereo, initial poses can still be estimated using P3P, but LBA equations are different.

4.4.1. Definition of reference frame

In order to deal with multi-camera case, a reference frame is defined on the vehicle body (*cf.* red in Fig. 6). Vehicle pose referenced to the world at time t is expressed as $(\boldsymbol{R}_{wv}^t, \boldsymbol{C}_{wv}^t)$. The rigid transformation from vehicle

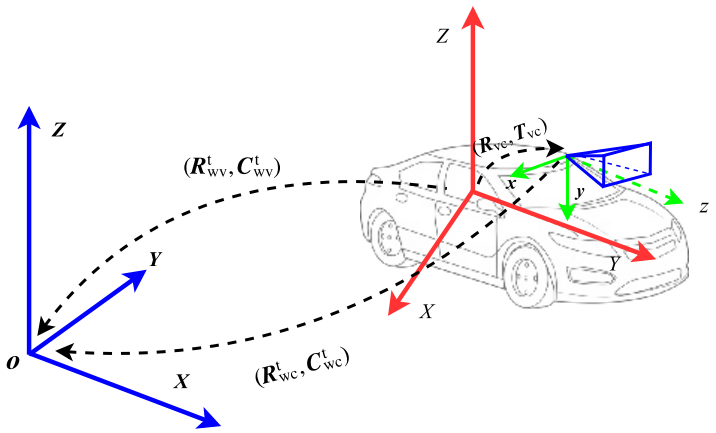


Figure 6: Transformation from camera to reference frame.

to camera is noted as $\boldsymbol{\Gamma}_{vc}$, it is composed of a rotation matrix $\boldsymbol{R}_{vc}$ and a translation vector $\boldsymbol{T}_{vc}$. These parameters can be calibrated beforehand, so the absolute pose of camera c can be estimated by:

$$\begin{cases} \boldsymbol{R}^t_{wc} = \boldsymbol{R}_{vc}\boldsymbol{R}^t_{wv} \\ \boldsymbol{C}^t_{wc} = \boldsymbol{R}^{\boldsymbol{t}}_{\boldsymbol{wv}}{}^{T}\boldsymbol{T}_{vc} + \boldsymbol{C}^t_{wv} \end{cases} \tag{15}$$

where, $\boldsymbol{R}^t_{wc}, \boldsymbol{C}^t_{wc}$ are absolute orientation and position for camera at time t.

4.4.2. LBA equations for multi-camera system

Considering Eq. 15, the rigid transformation can be integrated into Eq. 9, the back projection error for multi-camera system is computed by :

$$\boldsymbol{v}^c_t = F(\boldsymbol{P}^t_{wv}, \boldsymbol{\Gamma}_{vc}, \boldsymbol{X}_t) - \mathbf{x}^c_t, \tag{16}$$

where $\boldsymbol{v}^c_t$ are residuals for camera c. $\boldsymbol{X}_t$ represents tie points (expressed in world coordinate system). The $F(\boldsymbol{P}^t_{wv}, \boldsymbol{\Gamma}_{vc}, \boldsymbol{X}_t)$ is a back-projection function for tie points to every camera.

The uncertainties of rigid transformation parameters are taken into account in LBA. Prior values of Γ_{vc} are noted as Γ^0_{vc}. Combining error equations of prior poses, rigid transformation and back-projection errors, the joint error equations are:

$$\begin{cases} \boldsymbol{v}_p = \boldsymbol{P}^p_{wv} - \boldsymbol{P}^{\boldsymbol{p}}_{\boldsymbol{wv}}{}^{0} \\ \boldsymbol{v}_\Gamma = \boldsymbol{\Gamma}_{vc} - \boldsymbol{\Gamma}^0_{vc} \\ \boldsymbol{v}^c_t = F(\boldsymbol{P}^t_{wv}, \boldsymbol{\Gamma}_{vc}, \boldsymbol{X}_t) - \mathbf{x}^c_t \end{cases} \tag{17}$$

To obtain optimal estimates of the parameters, we minimize the cost function, written as:

$$[\hat{\boldsymbol{P}_{wv}^{p}}, \hat{\boldsymbol{P}_{wv}^{n}}, \hat{\boldsymbol{\Gamma}_{vc}}, \hat{\boldsymbol{X}_{t}}] = \underset{\boldsymbol{P}_p, \boldsymbol{P}_n, \boldsymbol{\Gamma}_c, \boldsymbol{X}_t}{\operatorname{argmin}} \{\frac{1}{2}(\boldsymbol{v_t}^T \boldsymbol{\Sigma}_t^{-1} \boldsymbol{v_t} + \boldsymbol{v}_p^T \boldsymbol{\Sigma}_p^{-1} \boldsymbol{v}_p + \boldsymbol{v}_\Gamma^T \boldsymbol{\Sigma}_\Gamma^{-1} \boldsymbol{v}_\Gamma)\} \tag{18}$$

where:

$[\hat{\boldsymbol{P}_{wv}^{p}}, \hat{\boldsymbol{P}_{wv}^{n}}, \hat{\boldsymbol{\Gamma}_{vc}}, \hat{\boldsymbol{X}_{t}}]$: optimal estimations of parameters.
$\boldsymbol{v}_\Gamma$: residuals of rigid transformation for all cameras
$\boldsymbol{\Sigma}_\Gamma$: covariance matrix of Γ_{vc}^0

In LBA, Eq. 17 is linearized with a first-order Taylor expansion. The covariance matrix is still computed from normal matrix based on Schur complement.

4.5. Variance Component Estimation

We assume that all feature points have the same precision, noted as σ_t. Thus, the covariance matrix for all interest points can be expressed as $\boldsymbol{\Sigma}_t = \sigma_t^2 \boldsymbol{I}$. The existing LBA methods regard σ_t as one pixel (Eudes and Lhuillier, 2009), but different interest point detector has different precision. For conventional BA, the variation of σ_t does not influence the estimates of parameters, because it only influences the scale of uncertainties. However, we propagate uncertainties over time, so we need a proper scale in LBA. In particular, incorrect scale for the uncertainties will lead to a wrong generation of search areas for matching, tracking and landmark registering which will be introduced later. The variance scale can be estimated using Variance Component Estimation (VCE) (Luxen, 2003). As LBA at first step is conventional BA. We set $\boldsymbol{\Sigma}_t$ as identity matrix at first. After BA optimization, the variance of interest points is computed via:

$$\hat{\sigma_t}^2 = \frac{\hat{\boldsymbol{v}_t}^T \hat{\boldsymbol{v}_t}}{r} \tag{19}$$

$\hat{v_t}$: residual vector after adjustment
r : the number of redundant observations.

Then the estimation of σ_t will be used for other LBA steps.

5. Integration of geo-referenced landmarks

In order to reduce the drift, geo-referenced landmarks are integrated into VO. There are three key steps for landmarks integration: 1) registering 3D landmarks into images; 2) generating a GCP and its measurement in an image from a 3D-2D correspondence; 3) considering the GCPs and their 2D measurements in LBA equations and resolving the constrained problem.

5.1. Geo-referenced landmarks

In this paper, we regard geo-referenced road signs and road marks as landmarks for localization. These two types of landmarks are reconstructed using the data acquired by MMSs. Each landmark is represented as a 3D polygon for which the normal vector is known. In addition, we know the category (e.g. prohibition, warning, obligation or road mark) of each object.

5.2. Selection of landmark candidates

Denote the pose of $\boldsymbol{I}_t$ as $\boldsymbol{P}_t^{'}$ estimated by VO. Fig. 7 illustrates our method for selection of 3D landmarks candidates for image $\boldsymbol{I}_t$. There are three steps:

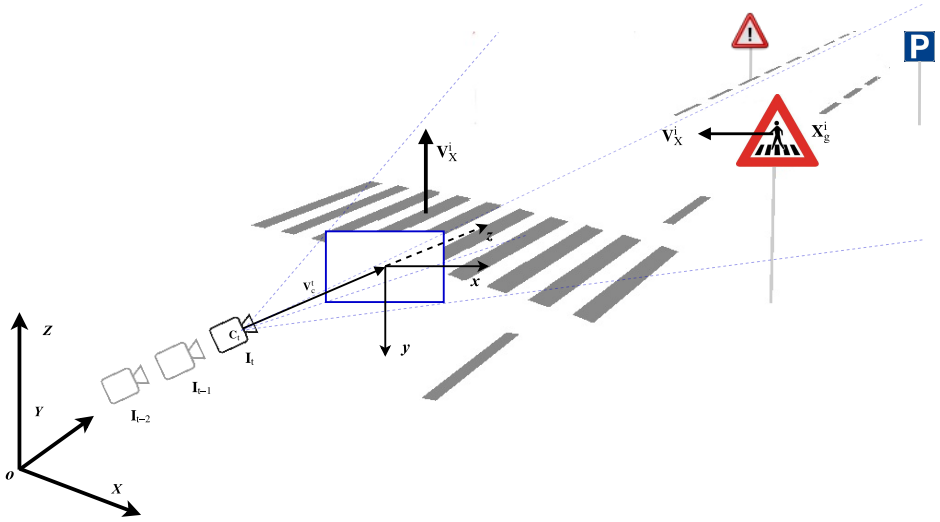


Figure 7: $\boldsymbol{C}_t$: absolute position of image t. $\boldsymbol{V}_c^t$: optical axis direction of camera. $\boldsymbol{V}_X^i$: the normal vector of landmark.

(1) Defining searching space. With a given threshold TH_s, the criterion for this step is:

$$\|\boldsymbol{X}_g^i - \boldsymbol{C}_t^{'}\| < TH_s,$$

where $\boldsymbol{X}_g^i$ is the position of 3D landmark center and $\boldsymbol{C}_t^{'}$ is the estimated position of camera.

(2) Rejecting invisible landmarks. Only landmarks inside image FOV are visible for the current image. The landmark is marked as visible when their back projection is inside image plane.

(3) Checking the normal direction of landmark. Some landmarks are inside the image FOV (*cf.* gray road signs in Fig. 7), but they are the signs for vehicles moving at the opposite direction. Note $\boldsymbol{V}_X^i$ as normal direction of the landmark which is determined by the plane equation of each 3D landmark. Then the intersection angle between $\boldsymbol{V}_X^i$ and $\boldsymbol{V}_c^t$ should not be smaller than 90°, that means:

$$\boldsymbol{V}_c^t \bullet \boldsymbol{V}_X^i \leq 0. \tag{20}$$

5.3. Uncertainty propagation for landmark registering

Now the 3D landmark candidates are registered into the images. In order to limit the scope for landmark registering, a coarse location can be estimated for each candidate. Then, uncertain regions are generated in an image to determine the search area.

A 3D vertex $\boldsymbol{X}_j$ of a landmark can be projected into image plane using pinhole projection function. The covariance matrix of image pose $\boldsymbol{\Sigma}_{\boldsymbol{P}_t}$ is estimated by LBA and the precision of 3D landmarks $\boldsymbol{\Sigma}_{\boldsymbol{X}_j}$ is known (obtained during reconstruction). Thus, the covariance matrix of $\boldsymbol{\Sigma}_{\mathbf{x}_j}$ can be estimated using uncertainty propagation principle for nonlinear equations:

$$\boldsymbol{\Sigma}_{\mathbf{x}_j} = \begin{bmatrix} \frac{\partial F}{\partial \boldsymbol{P}_t} & \frac{\partial F}{\partial \boldsymbol{X}_j} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{P}_t} & 0 \\ 0 & \boldsymbol{\Sigma}_{\boldsymbol{X}_j} \end{bmatrix} \begin{bmatrix} \frac{\partial F}{\partial \boldsymbol{P}_t} \\ \frac{\partial F}{\partial \boldsymbol{X}_j} \end{bmatrix} \tag{21}$$

The processing for multi-camera case is the same, but rigid transformation from camera to reference frame need to be considered. Note $\mathbf{x}_j^i$ as 2D location of a landmark in image captured by camera i. Its covariance matrix is:

$$\boldsymbol{\Sigma}_{\mathbf{x}_j^i} = \begin{bmatrix} \frac{\partial F}{\partial \boldsymbol{P}_t} & \frac{\partial F}{\partial \boldsymbol{\Gamma}_i} & \frac{\partial F}{\partial \boldsymbol{X}_j} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{P}_t} & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{\boldsymbol{\Gamma}_i} & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_{\boldsymbol{X}_j} \end{bmatrix} \begin{bmatrix} \frac{\partial F}{\partial \boldsymbol{P}_t} \\ \frac{\partial F}{\partial \boldsymbol{\Gamma}_i} \\ \frac{\partial F}{\partial \boldsymbol{X}_j} \end{bmatrix} \tag{22}$$

Then the uncertain region for each projection of landmark vertex can be calculated from the covariance matrix.

5.4. Registering a landmark in image

The searching area is the bounding box of all the uncertain regions for each vertex of a landmark candidate (*cf.* three vertexes for a triangular road sign in Fig. 8). Now the goal is to register a 3D landmark inside the searching area. Although both road signs and road marks are well-defined objects,

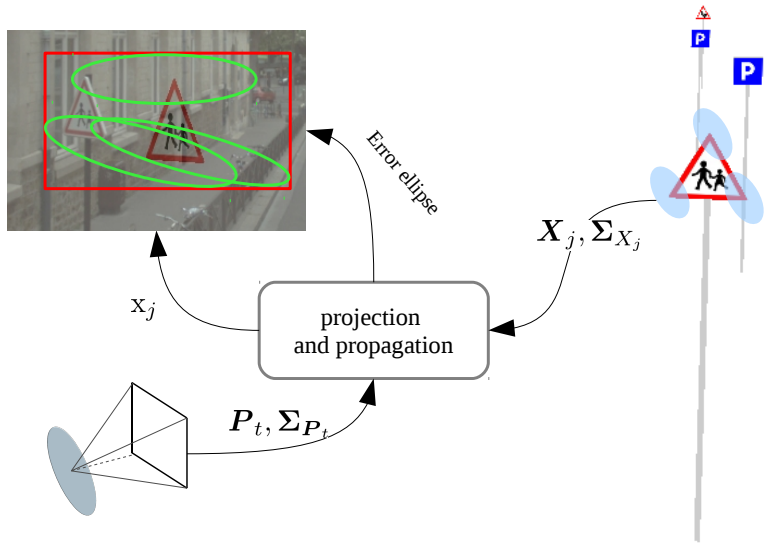


Figure 8: Generation of the searching area for a landmark in the image.

road signs have stronger visual and geometric properties than road marks. In particular, road signs are usually individual objects. These features of road signs enable us to recognize the signs, thus they are detected inside the searching area using the algorithm proposed by Soheilian et al. (2013). For road marks, the precise location of each one is obtained by optimizing an objective function in a regular MCMC (Markov Chain Monte Carlo) scheme (Soheilian et al., 2016). One successful landmark registration will generate one GCP, that is the center of a 3D landmark. The measurement of each GCP in an image is determined by the center of the registered landmark.

5.5. Integration of GCPs for monocular system

To deal with the GCPs, their back-projection errors are computed and minimized in LBA :

$$^{j}\boldsymbol{v}_c^i = F(\boldsymbol{P}_p, \boldsymbol{P}_n, \boldsymbol{X}_g^i) - {}^{j}\mathbf{x}_g^i, \tag{23}$$

where:

$\boldsymbol{X}_g^i$: GCP generated from i^{th} landmark.
$^{j}\mathbf{x}_g^i$: location of i^{th} landmark in j^{th}image.
$^{j}\boldsymbol{v}_c^i$: back projection errors of GCP $\boldsymbol{X}_g^i$ in image j

The form of Eq. 23 is the same as the back-projection error for normal tie points (*cf.* Eq. 11). However, $\boldsymbol{X}$ is a set of unknowns in Eq. 11, but $\boldsymbol{X}_g$ and its uncertainty are known for GCP. Denote $\boldsymbol{X}_g^0$ as the prior measurements of GCPs, determined by the centers of landmarks. $\sigma_X, \sigma_Y, \sigma_Z$ stand for the precision of $\boldsymbol{X}_g^0$. An extra error equation is added for each 3D GCP while its uncertainty is considered:

$$\boldsymbol{v}_g^i = \boldsymbol{X}_g^i - \boldsymbol{X}_g^{i\,0} \tag{24}$$

where, $\boldsymbol{v}_g^i$ represents the residuals of i^{th} GCP. Therefore, two types of error equations, expressed in Eq. 23 and 24, are added into LBA equations. The precision of landmarks in image is noted as σ_c, which is determined by the landmark registering algorithm. The covariance matrix for an image control point in image j is: $\boldsymbol{\Sigma}_c^j = \sigma_c^2 \boldsymbol{I}$ Meanwhile, the covariance matrix of every GCP $\boldsymbol{X}_g^{i\,0}$, is: $\boldsymbol{\Sigma}_g^i = diag(\sigma_X^2, \sigma_Y^2, \sigma_Z^2)$. Combining with Eq. 11 and 12, the new cost function can be written as:

$$[\hat{\boldsymbol{P}}_p, \hat{\boldsymbol{P}}_n, \hat{\boldsymbol{X}}_t, \hat{\boldsymbol{X}}_g] = \underset{\boldsymbol{P}_p, \boldsymbol{P}_n, \boldsymbol{X}_t, \boldsymbol{X}_g}{\operatorname{argmin}} \{\frac{1}{2}(\boldsymbol{v}_t^T \boldsymbol{\Sigma}_t^{-1} \boldsymbol{v}_t + \boldsymbol{v}_p^T \boldsymbol{\Sigma}_p^{-1} \boldsymbol{v}_p + \boldsymbol{v}_c^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{v}_c + \boldsymbol{v}_g^T \boldsymbol{\Sigma}_g^{-1} \boldsymbol{v}_g)\}, \tag{25}$$

where, $\hat{\boldsymbol{P}}_p, \hat{\boldsymbol{P}}_n, \hat{\boldsymbol{X}}_t, \hat{\boldsymbol{X}}_g$ are the optimal estimations of the parameters. $\boldsymbol{v}_c$ stands for all the back projection errors for GCPs, $\boldsymbol{v}_g$ means all GCPs residuals. $\boldsymbol{X}_g$ is a group of GCPs used in LBA window and $\mathbf{x}_g$ represents the locations of GCPs in images.

5.6. Integration of GCPs for multi-camera system

In multi-camera case, the landmark registering is performed for more images. The strategy about GCPs integration with LBA is the same as mono camera, but the rigid transformation from camera to reference frame is embedded into the equation system. Thus, the error equation of back projections for a GCP $\boldsymbol{X}_g^i$ in j^{th} image is given by:

$$^j\boldsymbol{v}_c^i = F(\boldsymbol{P}_p, \boldsymbol{P}_p, \boldsymbol{\Gamma}_i, \boldsymbol{X}_g^i) - {}^j\mathbf{x}_g^i.$$

Combining with Eq. 24 and the error equations for multi-camera introduced in section 4.4, the optimal estimates of the parameters can be obtained by minimizing the following cost function:

$$\begin{aligned}[\hat{\boldsymbol{P}}_p, \hat{\boldsymbol{P}}_n, \hat{\boldsymbol{X}}_t, \hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{X}}_g] = \underset{\boldsymbol{P}_p, \boldsymbol{P}_n, \boldsymbol{X}_t, \boldsymbol{\Gamma}, \boldsymbol{X}_g}{\operatorname{argmin}} \{\frac{1}{2}(\boldsymbol{v}_p^T \boldsymbol{\Sigma}_p^{-1} \boldsymbol{v}_p + \\ \boldsymbol{v}_\Gamma^T \boldsymbol{\Sigma}_\Gamma^{-1} \boldsymbol{v}_\Gamma + \boldsymbol{v}_t^T \boldsymbol{\Sigma}_t^{-1} \boldsymbol{v}_t + \boldsymbol{v}_c^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{v}_c + \boldsymbol{v}_g^T \boldsymbol{\Sigma}_g^{-1} \boldsymbol{v}_g)\}\end{aligned} \tag{26}$$

Nonlinear least squares method is applied to resolve the problem. The number of GCPs and corresponding image points is very small in comparison to normal tie points, so the required extra computation is negligible.

6. Experiments

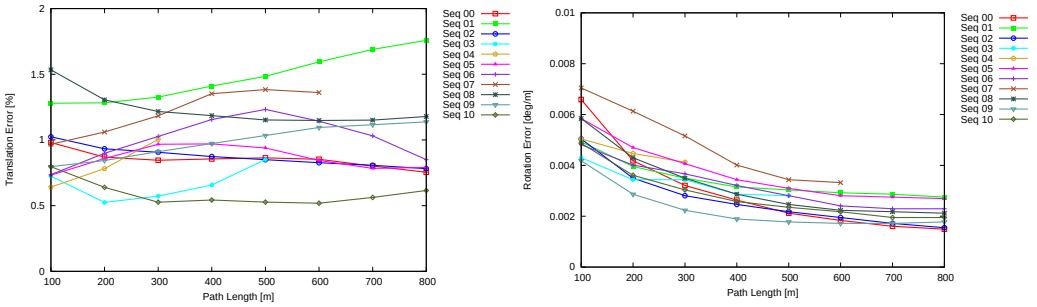
There are three parts in our experiments. First, our method is compared with the state-of-the-art approaches on KITTI benchmark (Geiger et al., 2012), that contains a set of conventional stereo sequences. Then, the performance of different camera configurations is tested for localization using our own data acquired by a MMS developed in IGN (Paparoditis et al., 2012). At last, the improvement of landmark integration for different camera configurations is demonstrated.

6.1. Evaluation of stereo based VO on KITTI benchmark

All images for VO evaluation on KITTI sites are rectified. Sequences 00-10 are used as training data and their ground truth poses are provided. The VO results from sequence 11 to 21 can be submitted to the KITTI website and compared with other methods. No geo-referenced landmarks are integrated into VO approach for KITTI datasets.

Although our method and the approaches like ORB-SLAM and LSD-SLAM can use monocular sequences for VO, the initialization strategy is

different. LSD-SLAM and ORB-SLAM sets a relative scale, while our VO method sets the absolute scale using a GPS receiver. For stereo sequences, the scale is set based on the length of baseline. Thus, we only compare the approach of stereo based VO. Relative measures (relative translation error t_{rel} and relative rotation error r_{rel}) are provided by KITTI site that computes the errors in fixed distances over sequence (Geiger et al., 2012), as shown in Fig. 9. The t_{rel} for each case is smaller than 2.0% at different distances using the proposed VO approach.



(a) t_{rel} (b) r_{rel}

Figure 9: Relative translation and rotation errors of our method using training data of KITTI.

The proposed VO approach is an extension of our previous work (Qu et al., 2015; Soheilian et al., 2016), which applied FLANN based matching strategy using SIFT features. In the new method, a propagation based matching and tracking method is proposed to improve both efficiency and accuracy (*cf.* Table. 1). Compared with the original matching method, the new approach reduces the translation error from 2.5% to 0.97% and the processing time for each frame is reduced from $1.2s$ to $0.17s$ for KITTI datasets.

Table 1: Improvement of propagation based matching and tracking method.

	$t_{rel}(\%)$	Runtime (s)
Previous approach	2.5	1.2
New method	**0.97**	**0.17**

We compare our method with ORB-SLAM2 (Mur-Artal and Tardós, 2016) and S-LSD-SLAM (stereo LSD-SLAM) (Engel et al., 2015), shown in Table 2.

Table 2: Relative localization errors for our method, ORB-SLAM2 and S-LSD-SLAM.

	ORB-SLAM2 t_{rel} (%)	S-LSD-SLAM t_{rel}(%)	Our method t_{rel}(%)
KITTI_00	0.70	**0.63**	0.86
KITTI_01	**1.39**	2.36	1.45
KITTI_02	**0.76**	0.79	0.88
KITTI_03	0.71	1.01	**0.64**
KITTI_04	**0.48**	0.38	0.75
KITTI_05	**0.40**	0.64	0.86
KITTI_06	**0.51**	0.71	1.00
KITTI_07	**0.50**	0.56	1.14
KITTI_08	**1.05**	1.11	1.24
KITTI_09	**0.87**	1.14	0.97
KITTI_10	**0.60**	0.72	0.61
Average	**0.72**	0 .917	0.97

Both ORB-SLAM2 and S-LSD-SLAM detect loops and apply global bundle adjustment for optimization if the loop is successfully detected. In KITTI datasets, sequences 00, 02, 05, 06 and 07 contain loops, so the localization errors for both ORB-SLAM2 and S-LSD-SLAM are smaller than ours, because no loop closure is conducted in our approach. For the other sequences, our method achieves better accuracy than S-LSD-SLAM most of the time (*cf.* sequence 01, 03 ,09, 10) and slightly worse than ORB-SLAM2.

The results for stereo sequences from 11 to 21 were submitted to KITTI website under the name of SLUP[1] (Stereo based Localization considering Uncertainty Propagation). The average translation error is 1.25% and the average rotation error is 0.0041 deg/m. Compared with training datasets, the accuracy of testing data is decreased. By analyzing the sequences, we found that the sequences 12, 20 and 21 were captured on a highway which can produce inaccurate pose estimation because of repeatable textures and unreliable matches on moving vehicles in these three sequences. In training datasets, sequence 01 had the similar situation, so its localization errors is bigger than other sequences (*cf.* Fig 9).

The top stereo based VO achieve 0.66% for translation and 0.0015 deg/m

[1]http://www.cvlibs.net/datasets/kitti/eval_odometry.php

for rotation, known as *SOFT2*. Table 3 presents the performance of some state-of-the-art VO methods on KITTI benchmark suite. For the published

Table 3: Performance of state-of-the-art VO methods.

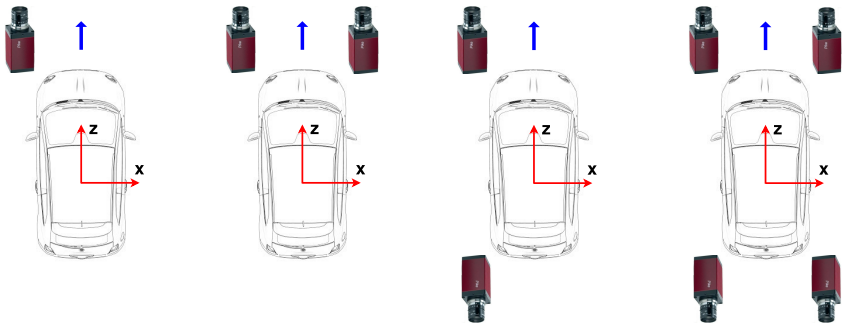
Method	Rank	t_{rel}	r_{rel} (deg/m)	Time	Loop closure
RotRocc++	5	0.83 %	0.0026	0.26 s	No
SOFT	8	0.88 %	0.0022	0.1 s	No
ORB-SLAM2	17	1.15 %	0.0027	0.06 s	Yes
S-PTAM	22	1.19 %	0.0025	0.03 s	Yes
S-LSD-SLAM	23	1.20 %	0.0033	0.07 s	Yes
SLUP	**27**	**1.25** %	**0.0041**	**0.17 s**	**No**

work, the best rank is five for RotRocc++ (Buczko and Willert, 2016), which filters the outliers by normalizing the back-projection errors. SOFT (Cvišić and Petrović, 2015) achieved the similar accuracy and they take strategies to select the most robust tracks for pose estimation. These two approaches give us some new ideas to improve the accuracy in our future work. Compared with the VO methods (S-PTAM (Pire et al., 2016), S-LSD-SLAM (Engel et al., 2015) and ORB-SLAM2 (Mur-Artal and Tardós, 2016)), our method achieves the similar accuracy.

In the training sequences (00-10), ORB-SLAM2 performs much better than ours (0.72% VS 0.97 %), but difference is less significant in test data (1.15% VS 1.25 %). Our method is designed to reduce the drifts of VO by integrating geo-referenced landmarks and doesnt perform loop closure. This can explain the difference of our score and those of ORB-SLAM2 and S-LSD-SLAM (both using loop-closure technique) on KITTI benchmark that doesnt provide any geo-referenced landmark. Although the localization error of stereo based VO can be less than 1 % with the state-of-the-art methods, the absolute errors still increase over time, it might be over several meters after a long trajectory, which is not sufficient for precise localization.

6.2. Evaluation of different camera configuration

Multiple cameras are mounted on our MMS, hence different camera configuration can be composed for localization. In this experiment, four camera configurations are tested (*cf.* Fig. 10). The focal length is 10 mm for each camera and the image size is 1920×1024 pixels. The FOV of each camera



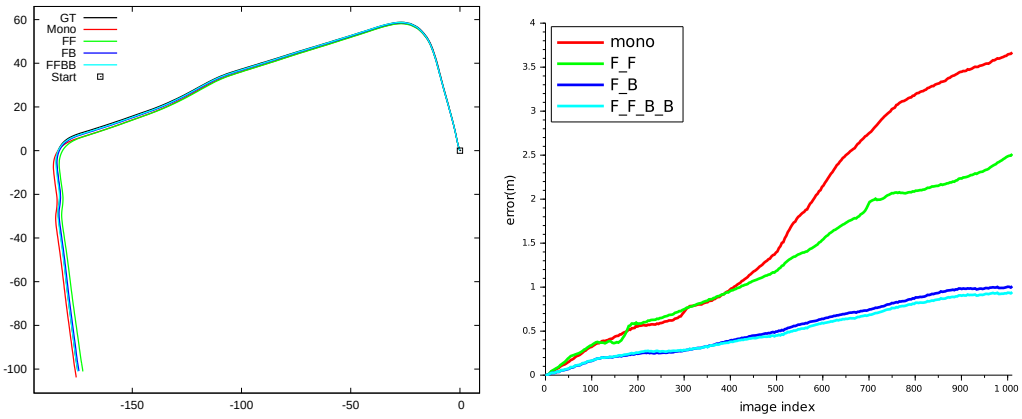
(a) Mono (b) F_F (c) F_B (d) F_F_B_B

Figure 10: Design of camera configuration. **F**: Forward looking. **B**: Backward looking. F_F: conventional stereo. F_B: non-overlap stereo. F_F_B_B: multi-camera.

is 70° in horizontal and 42° in vertical. All cameras were calibrated beforehand. A precise GPS/INS/odometer navigation system was used to provide the ground truth with absolute localization accuracy of 10 cm. 1100 frames over $340m$ trajectory are captured by each camera.

The VO results from different camera configurations were compared with ground truth in our experiments. The start point is provided by a GPS receiver and the absolute scale is determined by the distance from the first frame to second key frame measured by GPS. The estimated paths, compared with ground truth are demonstrated in Fig. 11a. The absolute errors which are the Euclidean distances from estimated positions to ground truth, are calculated, as shown in Fig. 11b. From this diagram, we can learn that F_B and F_F_B_B perform better than mono and F_F, because they have larger FOV. Therefore, we can compensate the drift with larger FOV when the angular resolution of the image is same. Furthermore, the maximum, mean errors and RMSE for different camera configuration are calculated (*cf.* Table 4). Together with the curves shown in Fig. 11b, they reveal the superiority of rear-front looking configurations (F_B and F_F_B_B) compared with exclusively front looking ones (Mono and F_F). The drift is reduced by four times from Mono to F_B. The gain obtained by F_F_B_B comparing to F_B is less significant.

In practice, the design of camera configuration (number and position of the embedded cameras) should take into account the gain of precision but also the cost of sensors, energy consumption and also computation time. Meanwhile, the absolute error of localization is 0.53m (RMSE) over 340m



(a) Estimation of paths

(b) Absolute errors.

Figure 11: Estimation of locations for different camera configurations.

Table 4: Accuracy of localization for different camera configurations over a trajectory of $340m$.

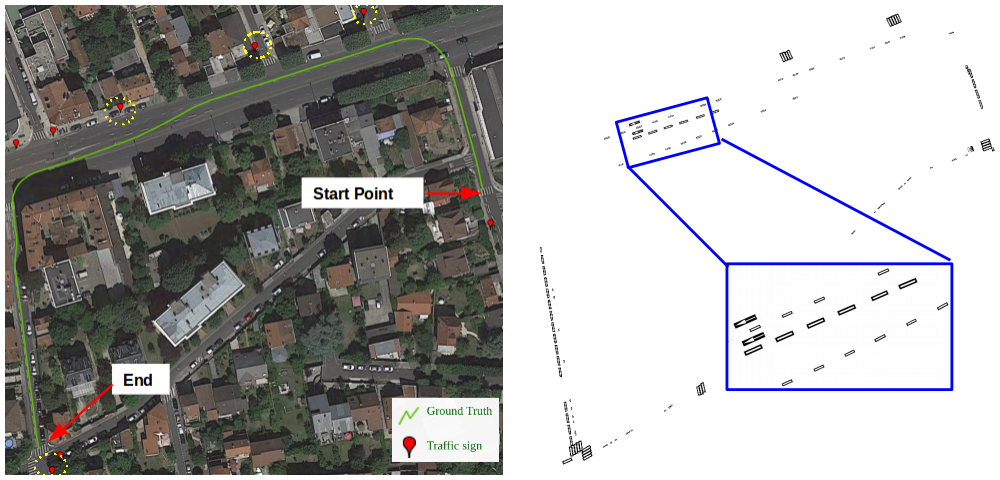
	Maximum (m)	Mean (m)	RMSE (m)
Mono	3.67	2.75	2.13
F_F	2.79	1.65	1.43
F_B	1.26	0.68	0.59
F_F_B_B	**0.99**	**0.61**	**0.53**

trajectory even using four camera configuration and the RMSE will increase to a larger value after moving a longer distance. Therefore, the integration of geo-referenced landmarks is needed for precise localization.

6.3. Evaluation of VO integrated with geo-referenced landmarks

Fig. 12 shows the ground truth (same datasets as previous section) and landmark maps. There are eight geo-referenced road signs (*cf.* red dots in Fig. 12a) in test area, reconstructed from a set of images captured by a mobile mapping system. Fig. 12b demonstrates the geo-referenced road marks used in this experiment. The precisions of landmarks are better than 10cm.

For the same test site, four different experiments are conducted. First, we test our VO method using multiple camera configurations, introduced in section 6.2. Second, only geo-referenced road signs (RS) are integrated



(a) Traffic sign locations (b) Geo-referenced road marks

Figure 12: Trajectory of ground truth and landmark maps.

into the VO. Third, only geo-referenced road marks are used. Finally, both road signs and road marks (RS+RM) are considered into our system for localization.

Table 5: Relative translation error using different camera configurations and different types of landmarks (%).

	Mono	F_F	F_B	F_F_B_B
VO	1.12	0.88	0.40	0.32
VO + RS	0.68	0.44	0.29	0.12
VO + RM	0.030	**0.012**	**0.029**	**0.002**
VO+RS+RM	**0.029**	0.013	**0.029**	**0.002**

It is straightforward that the accuracy is improved with the integration of landmarks. Only four out of eight road signs are successfully matched in this experiment, marked with yellow circles in Fig. 12a. However, the accuracy is compensated significantly, for instance the translation error is reduced down to 0.68% for mono (1.12% without landmarks, cf. Table 5). For the case of road marks integration, the accuracy is close to the performance of VO+RS+RM, because there are 73 geo-referenced road marks that are successfully integrated with VO approach while only four road signs are used.

Table 6 contains the maximum absolute errors, mean errors and RMSE

for different configurations. If the landmarks of RS+RM are integrated, the maximum absolute errors for all camera configurations are less than $0.31m$ and the RMSE of localization is less than $0.17m$. The more landmarks are registered, the better accuracy is achieved. However, few landmarks can still improve the accuracy significantly. With larger FOV camera configuration such as F_B, the RMSE of localization is limited to 0.52m over 340 trajectory with the integration of only four road signs (cf. Table 6).

Table 6: Results of localization with different landmarks.

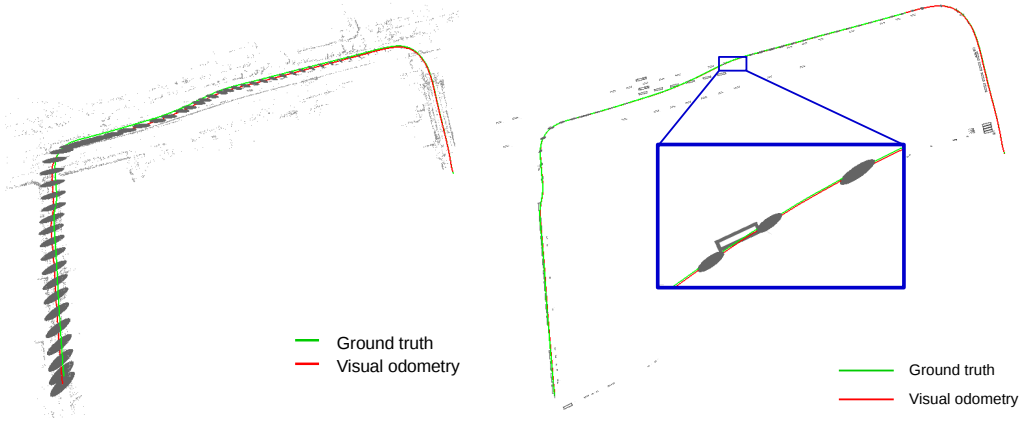
	VO+RS (m)			VO+RM (m)			VO+RS+RM (m)		
	Max	Mean	RMSE	Max	Mean	RMSE	Max	Mean	RMSE
Mono	2.03	0.93	1.16	0.32	0.16	0.17	0.31	0.16	0.17
F_F	1.38	0.73	0.62	0.28	0.12	0.13	0.24	0.12	0.13
F_B	0.76	0.47	0.52	0.25	0.12	0.13	0.24	0.11	0.13
F_F_B_B	0.53	0.33	0.36	0.23	0.12	0.12	**0.23**	**0.11**	**0.12**

The error ellipsoids of localization with and without the integration of landmarks for monocular VO are shown in Fig. 13. The confident level is set as 99 %. In Fig. 13a, we can see the growing of trajectory's drift over time, but the drift can be overcome when the geo-referenced landmarks are integrated. The localization is very precise as the estimated path is very close to the ground truth in Fig. 13b. Meanwhile, the uncertainty of localization is also reduced, the error ellipsoids are always small over sequence as shown in Fig. 13b. For better visualization, the size of error ellipsoids shown in Fig. 13 exaggerated five times. We set the same scale for both cases (with and without integration of landmarks) in order to make the comparison easier.

The extra computation caused by the integration of GCPs for LBA can be ignored because the number of GCPs is far smaller than the number of normal tie points. In the current system, road sign detection is efficient which only needs approximately $0.15s$ per object in our experiment. However, the road mark matching is off-line processing at the current time, which is based on MCMC. It takes $10 - 40s$ per object due to a high number of iterations.

7. Conclusion

This paper proposed a cost effective and precise localization method. The system uses one or more optical cameras for ego-motion estimation and a set of geo-referenced landmarks in order to reduce the drift. It requires an approximate initialization that can be provided by a low-cost GNSS receiver.



(a) Without landmarks (b) Integration of landmarks

Figure 13: Error ellipsoids for localization using mono camera.

This paper is an extension of our previous work (Qu et al., 2015; Soheilian et al., 2016). Three main contributions were introduced in this paper. First, a propagation based matching and tracking method was proposed to improve both efficiency and accuracy of interest points matching and tracking. The proposed VO was evaluated on KITTI benchmark and compared with the state of the art approaches. The accuracy of proposed VO approach is close to the methods such as ORB-SLAM2, S-LSD-SLAM and S-PTAM, but we should mention that the runtime is longer than these methods. The main contribution of our method is to approach the performances of the state of the art in VO while enabling integration of geo-referenced landmark for reducing the drift

Second, the VO method was extended to handle any camera configuration. The performance of four different configurations was evaluated using real data. It divulged that front-rear looking camera configuration performs better than exclusively front looking stereo. Compared to mono camera configuration, using four cameras (two forward looking and two rear looking) reduces the drift up to four times.

Third, evaluation of the method using two different types of landmarks namely road signs and road markings on the same trajectory revealed that the drift of VO can be reduced up to 1.5-3 times by using a few road signs and up to 30-100 times by taking benefit from high density of road markings.

In future work, we will evaluate the robustness of the method to errors

of the map such as disappeared or moved objects and will investigate map updating using the users' sensors (those embedded on vehicles using the map for localization).

References

Agrawal, M., Konolige, K., 2006. Real-time localization in outdoor environments using stereo vision and inexpensive gps, in: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, IEEE. pp. 1063–1068.

Alonso, I.P., Llorca, D.F., Gavilán, M., Pardo, S.Á., García-Garrido, M.Á., Vlacic, L., Sotelo, M.Á., 2012. Accurate global localization using visual odometry and digital maps on urban environments. Intelligent Transportation Systems, IEEE Transactions on 13, 1535–1545.

Arth, C., Pirchheim, C., Lepetit, V., Ventura, J., 2015. Global 6dof pose estimation from untextured 2d city models. arXiv preprint arXiv:1503.02675 .

Bodensteiner, C., Hübner, W., Jüngling, K., Solbrig, P., Arens, M., 2011. Monocular camera trajectory optimization using lidar data, in: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, IEEE. pp. 2018–2025.

Brubaker, M., Geiger, A., Urtasun, R., 2013. Lost! leveraging the crowd for probabilistic visual self-localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3057–3064.

Buczko, M., Willert, V., 2016. Flow-decoupled normalized reprojection error for visual odometry, in: 19th IEEE Intelligent Transportation Systems Conference (ITSC).

Caron, G., Dame, A., Marchand, E., 2014. Direct model based visual tracking and pose estimation using mutual information. Image and Vision Computing 32, 54 – 63.

Cvišić, I., Petrović, I., 2015. Stereo odometry based on careful feature selection and tracking, in: Mobile Robots (ECMR), 2015 European Conference on, IEEE. pp. 1–6.

Davison, 2003. Real-time simultaneous localisation and mapping with a single camera, in: IEEE International Conference on Computer Vision (ICCV),, p. 1403.

Engel, J., Koltun, V., Cremers, D., 2016. Direct sparse odometry, in: arXiv:1607.02565.

Engel, J., Schöps, T., Cremers, D., 2014. LSD-SLAM: Large-scale direct monocular SLAM, in: European Conference on Computer Vision (ECCV).

Engel, J., Stckler, J., Cremers, D., 2015. Large-scale direct slam with stereo cameras, in: Ieee/rsj International Conference on Intelligent Robots and Systems, pp. 1935–1942.

Eudes, A., Lhuillier, M., 2009. Error propagations for local bundle adjustment, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 2411–2418.

Forster, C., Pizzoli, M., Scaramuzza, D., 2014. Svo: Fast semi-direct monocular visual odometry, in: IEEE International Conference on Robotics and Automation, pp. 15–22.

Furgale, P., Barfoot, T., 2010. Stereo mapping and localization for long-range path following on rough terrain, in: Robotics and Automation (ICRA), 2010 IEEE International Conference on, IEEE. pp. 4410–4416.

Galvez-Lpez, D., Tardos, J.D., 2012. Bags of binary words for fast place recognition in image sequences. IEEE Transactions on Robotics 28, 1188–1197.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite, in: Conference on Computer Vision and PatternRecognition (CVPR).

Gupta, A., Chang, H., Yilmaz, A., 2016. Gps-denied geo-localisation using visual odometry. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences III-3, 263–270.

Hervieu, A., Soheilian, B., Bredif, M., 2015. road marking extraction using model&data-driven RJ-MCMC. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences II-3/W4, 47–54.

Ji, S., Yuan, X., 2016. A generic probabilistic model and a hierarchical solution for sensor localization in noisy and restricted conditions. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences 41.

Klein, G., Murray, D., 2009. Parallel tracking and mapping on a camera phone, in: IEEE International Symposium on Mixed and Augmented Reality, pp. 83–86.

Kneip, L., Scaramuzza, D., Siegwart, R., 2011. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE. pp. 2969–2976.

Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., Furgale, P., 2015. Keyframe-based visual-inertial odometry using nonlinear optimization. International Journal of Robotics Research 34, 314–334.

Lhuillier, M., 2012. Incremental fusion of structure-from-motion and gps using constrained bundle adjustments. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34, 2489–2495.

Lothe, P., Bourgeois, S., Dekeyser, F., Royer, E., Dhome, M., 2009. Towards geographical referencing of monocular slam reconstruction using 3d city models: Application to real-time accurate vision-based localization, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 2882–2889.

Luxen, M., 2003. Variance component estimation in performance characteristics applied to feature extraction procedures, in: Pattern Recognition. Springer, pp. 498–506.

Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., Sayd, P., 2006. Real time localization and 3d reconstruction, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, IEEE. pp. 363–370.

Mur-Artal, R., Tardós, J.D., 2016. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. CoRR abs/1610.06475. URL: `http://arxiv.org/abs/1610.06475`.

Nistér, D., Naroditsky, O., Bergen, J., 2004. Visual odometry, in: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, IEEE. pp. I–652.

Paparoditis, N., Papelard, J.P., Cannelle, B., Devaux, A., Soheilian, B., David, N., HOUZAY, E., 2012. Stereopolis ii: A multi-purpose and multi-sensor 3d mobile mapping system for street visualisation and 3d metrology. Revue française de photogrammétrie et de télédétection , 69–79.

Pink, O., 2008. Visual map matching and localization using a global feature map, in: Computer vision and pattern recognition workshops, pp. 1–7.

Pire, T., Fischer, T., Civera, J., Cristoforis, P.D., 2016. Stereo parallel tracking and mapping for robot localization, in: Ieee/rsj International Conference on Intelligent Robots and Systems, pp. 1373–1378.

Qu, X., Soheilian, B., Paparoditis, N., 2015. Vehicle localization using mono-camera and geo-referenced traffic signs, in: IEEE Intelligent Vehicles Symposium (IV2015), Seoul, South Korea.

Quan, L., Lan, Z., 1999. Linear n-point camera pose determination. Pattern Analysis and Machine Intelligence, IEEE Transactions on 21, 774–780.

Rosten, E., Drummond, T., 2006. Machine learning for high-speed corner detection, in: Computer Vision–ECCV 2006. Springer, pp. 430–443.

Royer, E., Lhuillier, M., Dhome, M., Lavest, J.M., 2007. Monocular vision for mobile robot localization and autonomous navigation. International Journal of Computer Vision 74, 237–260.

Schreiber, M., Knoppel, C., Franke, U., 2013. Laneloc: Lane marking based localization using highly accurate maps, in: Intelligent Vehicles Symposium (IV), 2013 IEEE, IEEE. pp. 449–454.

Soheilian, B., Paparoditis, N., Boldo, D., 2010. 3d road marking reconstruction from street-level calibrated stereo pairs. ISPRS Journal of Photogrammetry and Remote Sensing 65, 347–359.

Soheilian, B., Paparoditis, N., Vallet, B., 2013. Detection and 3d reconstruction of traffic signs from multiple view color images. ISPRS Journal of Photogrammetry and Remote Sensing 77, 1–20.

Soheilian, B., Qu, X., Brdif, M., 2016. Landmark based localization: Lba refinement using mcmc-optimized projections of rjmcmc-extracted road marks, in: 2016 IEEE Intelligent Vehicles Symposium (IV), pp. 940–947. doi:10.1109/IVS.2016.7535501.

Thrun, S., Burgard, W., Fox, D., 2005. Probabilistic robotics. MIT press.

Tournaire, O., Soheilian, B., Paparoditis, N., 2006. Towards a sub-decimetric georeferencing of ground-based mobile mapping systems in urban areas: matching ground-based and aerial-based imagery using roadmarks, in: Proc. of the ISPRS Commission I Symposium, Marne-la-Vall ee, France. Interne.

Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W., 2000. Bundle adjustmenta modern synthesis, in: Vision algorithms: theory and practice. Springer, pp. 298–372.

Wei, L., Cappelle, C., Ruichek, Y., Zann, F., 2011. Gps and stereovision-based visual odometry: application to urban scene mapping and intelligent vehicle localization. International Journal of Vehicular Technology 2011.

Wei, L., Soheilian, B., Gouet-Brunet, V., 2014. Augmenting vehicle localization accuracy with cameras and 3d road infrastructure database, in: Computer Vision-ECCV 2014 Workshops, Springer. pp. 194–208.

Zhang, Z., Rebecq, H., Forster, C., Scaramuzza, D., 2016. Benefit of large field-of-view cameras for visual odometry, in: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 801–808. doi:10.1109/ICRA.2016.7487210.