Hindawi Publishing Corporation
International Journal of Vehicular Technology
Volume 2011, Article ID 439074, 17 pages
doi:10.1155/2011/439074

Research Article

GPS and Stereovision-Based Visual Odometry: Application to Urban Scene Mapping and Intelligent Vehicle Localization

Lijun Wei, Cindy Cappelle, Yassine Ruichek, and Frédérick Zann

Laboratoire Systèmes et Transports, Université de Technologie de Belfort-Montbéliard, 90010 Belfort, France

Correspondence should be addressed to Lijun Wei, lijun.wei@utbm.fr

Received 15 November 2010; Accepted 8 March 2011

Academic Editor: Hwangjun Song

We propose an approach for vehicle localization in dense urban environments using a stereoscopic system and a GPS sensor. Stereoscopic system is used to capture the stereo video flow, to recover the environments, and to estimate the vehicle motion based on feature detection, matching, and triangulation from every image pair. A relative depth constraint is applied to eliminate the tracking couples which are inconsistent with the vehicle ego-motion. Then the optimal rotation and translation between the current and the reference frames are computed using an RANSAC based minimization method. Meanwhile, GPS positions are obtained by an on-board GPS receiver and periodically used to adjust the vehicle orientations and positions estimated by stereovision. The proposed method is tested with two real sequences obtained by a GEM vehicle equipped with a stereoscopic system and a RTK-GPS receiver. The results show that the vision/GPS integrated trajectory can fit the ground truth better than the vision-only method, especially for the vehicle orientation. And vice-versa, the stereovision-based motion estimation method can correct the GPS signal failures (e.g., GPS jumps) due to multipath problem or other noises.

1. Introduction

Autonomous vehicle navigation and unmanned driving (e.g., localization, path planning, obstacle avoidance, etc.) have became important research areas in recent years. How to accurately localize the vehicle is one of the key issues for all these functions. Satellites-based navigation systems (e.g., GPS and Galileo) have been the most popular tools for outdoor vehicle global localization and navigation. It is a kind of absolute localization strategies that calculates the traveling distances of satellite signals from at least four visible satellites to on-board receivers, then uses trilateration method to compute the position of receivers. They can provide accurate positions in the long term, but as the signals are affected by atmospheric conditions, satellite positions, radio signal noises, and so forth, the accuracy in the short term is only to a few meters. Although the Real-Time Kinematic GPS (RTK-GPS) can deliver position up to centimeter accuracy, it does not work well anymore in some particular dense urban environments (e.g., urban canyons), as the satellite signals might be blocked or reflected by tall buildings. Insufficient satellites number or the multipath problem will decrease the availability on the position accuracy. Furthermore, the nonstationary noise of GPS might affect the GPS observation model.

In order to compensate the GPS problems, other types of methods are often added to provide accurate and robust vehicle localization. It permits to obtain the current position of vehicle based on its relative motion from the previous position. For example, the wheel encoder-based odometry can localize the vehicle by measuring the traveling distance and the elementary rotation. Nevertheless, wheel odometry-based localization suffers from wheel slippage in rock areas or muddy areas or from bad wheel radius estimation. Another kind of dead-reckoning-based localization approach is IMU (Inertial Measurement Unit). It achieves the purpose by measuring the 3D accelerations and 3D orientations of vehicle at every instant. Computer vision-based visual odometry methods have also been used in a lot of projects, even on Mars [1] as dead-reckoning localization method. Using laser or camera can effectively help to percept the environment, then assist vehicle localization and navigation by continuously mapping the real world and estimating the

relative vehicle motion. In outdoor environment, vehicle dynamic computation and scenery modeling make the use of vision method very challenging, especially for large-scale cases. Suddenly appearing pedestrians or moving vehicles will generate points with relative motion from the vehicle. Natural landmarks in areas with amount of repeating textures (e.g., trees, fences) also make the use of vision complicated. All these dead-reckoning methods can provide good accuracy in the short term; however, the trajectory will drift in the long term as errors accumulate from point to point.

In the absence of an ideal sensor for localization, an efficient solution is to integrate the absolute localization sensor GPS and other sensors together, and to take advantages of the best characteristics of every sensor. As in dense urban environments, there are amounts of visual landmarks, the vision-based method can supply both localization and mapping information. Cameras are often employed because they are less expensive and lighter than laser scanning system. In this paper, we integrate a stereoscopic system and a RTK-GPS sensor together. Stereovision is used to perform 3D reconstruction and to estimate camera motion. Stereovision method is adopted in our system because the baseline between left and right cameras is already known by calibration, thus the scale of Euclidean reconstruction is directly provided. At the same time, the precision of GPS position measurement is checked at a regular interval. Accurate GPS positions are used to adjust the vision-based vehicle trajectory and 3D maps. When the GPS signals cannot be received or only get signals without high precision, only vision-based method is used. Considering the errors and uncertainties during the whole process, several 2D and 3D outlier rejection strategies are adopted for the motion estimation, such as a relative depth constraint for the tracking step.

This paper is organized as follows. Section 2 makes a review of the vision and multisensor-based localization methods. Section 3 introduces the proposed method. Section 4 presents the used stereoscopic system, 3D landmark reconstruction, feature tracking, and robust outlier rejection mechanism. Section 5 details the camera egomotion estimation by stereovision-only method and the integration with accurate GPS positions. Section 6 presents the proposed mapping and localization results with the data acquired by a real electrical vehicle equipped with a stereoscopic sensor and a RTK-GPS receiver. Finally some conclusions and future perspectives are presented in Section 7.

2. Related Work

As our system is largely relying on the computer vision system, we will firstly make a review of two kinds of computer vision-based localization methods depending on whether the environment 3D model is known or not (with predefined model or simultaneously constructed model). Then, we concentrate on the stereovision-based visual odometry method. Finally, some existed localization methods using multisensor integration are presented.

The first vision-based localization approach is with a predefined model. It is realized by manually driving the vehicle (or mobile robot) along a desired path as the learning phase, then building an accurate 3D map of the environment (together with the corresponding camera motion) from learning sequence. After that, this model is used to locate and navigate the vehicle in real time. Such as in [2, 3], a single forward-looking camera with no calibration is used to build the model by structure from motion method. But as the baseline between two instants is unknown, the scale factor of reconstruction is ambiguous and should be estimated by entering the path length of the GPS trajectory as the last step. In [4], Lothe et al. propose to use a coarse 3D model of the environment (like GIS database), bundle adjustment, and some geometric constraints to align the reconstruction scale. However, if more than one camera are provided, the scale and scene geometry of the environment can be directly recovered through triangulation of the 3D points, such as Se et al. [5] and Nogueira et al. [6] use binocular vision method to build the 3D model. For both two methods, a final bundle adjustment (with Levenberg-Marquardt algorithm) is inevitable in order to refine the 3D coordinates and camera positions. And for the localization part, tracking the predefined model and searching for the corresponding key frames are still time-consuming and troublesome.

The second approach is the well-known simultaneous localization and mapping method (SLAM). It is a technique that can map the environment and localize the vehicle at the same time [7]. And it can be divided into two kinds of methods.

(1) The classic SLAM is with a camera motion model and based on probabilistic method using filtering [8]. The prediction of mobile motion is firstly obtained by motion model/vehicle dynamics, then the mean and covariance matrices of state vector (composed of vehicle and 3D landmarks positions) are corrected by observed natural or man-made landmarks.

(2) The vision-based SLAM can be separated as visual SLAM and visual odometry. Visual SLAM estimates the vehicle trajectory by matching features between a live map of the scene structure and the current image [9]. Visual odometry was first proposed by Nistér et al. in [10]. It estimates the path of cameras by continuously calculating the relative egomotion between images from the video flow, without any prior knowledge of the scene or a predefined motion model. We can also call it egomotion estimation. Features are detected and matched (or tracked) between pairs of frames, then the estimates of camera motion are produced from the corresponding features using 3D/2D [10] or 3D/3D [11] methods. Both monocular and stereoscopic systems are adopted to the ground vehicle localization: monocular omnidirectional camera which can give 360 degrees panoramic view of the surroundings is used in [12, 13] for wider range of view and better depth precision. Binocular vision [14–18] is used for the advantages of reconstruction. Some efforts are employed to reject outliers in the matching and tracking steps [19, 20]. Multicameras rig are also considered [21–23] to avoid 3D motion recovery ambiguity due to the small field of view and small depth variation in the field of view.

As there is no global optimization for the second vision-based method, the localization accuracy decreases as the

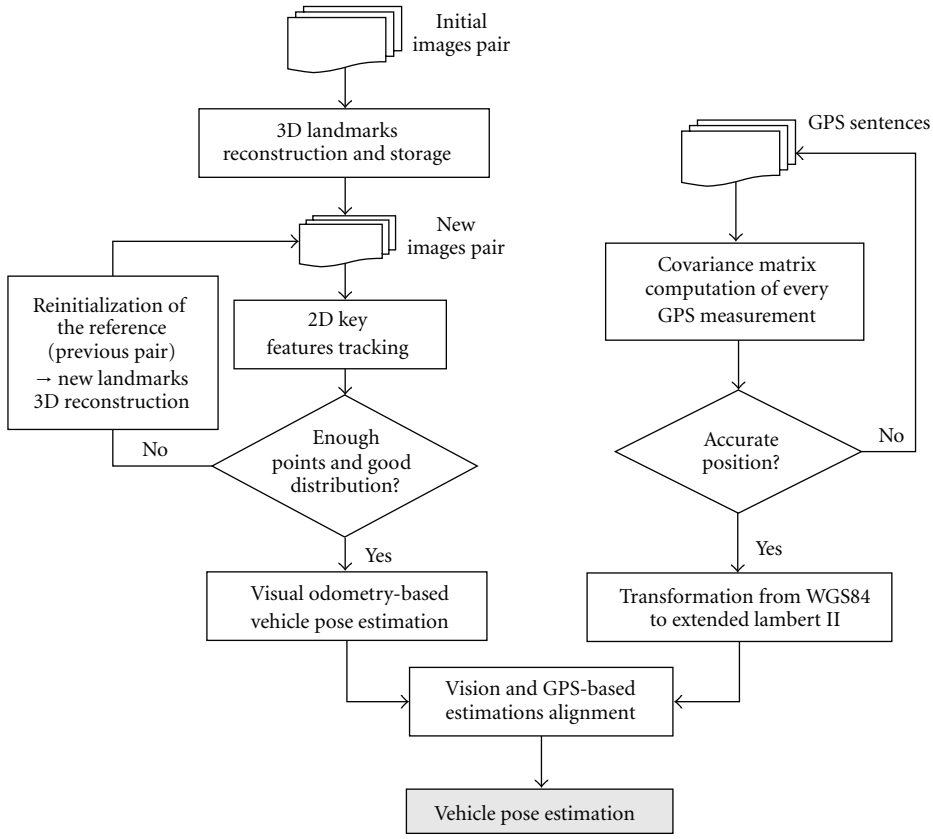


Figure 1: Overview of the proposed localization method.

traveling distance increases. Various methods (e.g., loop-closure) or the information from different sensor sources (e.g., GPS, inertia sensors, wheel encoders) are combined together to increase the precision of localization, such as El Najjar and Bonnifait integrate GPS with odometry [24]; Sukkarieh et al. combine GPS with IMU [25]; Cappelle et al. localize the vehicle by fusion of GPS, vision, and 3D GIS model [26]; Grimes and LeCun augment wheel odometry with visual orientation tracking to yield better localization accuracy [27].

Therefore, in our proposed approach, GPS and stereoscopic system are integrated together. Stereovision-based visual odometry method is used to estimate the vehicle egomotion at every instant, and the absolute localization GPS sensor is applied as the measurements to adjust the vehicle trajectory.

3. Methodology

Our proposed method is composed of three principal parts: 3D landmarks reconstruction, vision-based vehicle trajectory estimation, and motion-adjusting with accurate GPS positions. The flowchart is shown in Figure 1.

First of all, based on a video sequence from a stereoscopic system, image features are detected (Section 4.2) and matched (Section 4.3) for the initial stereo pair, then reconstructed into 3D space and stored as 3D landmarks after outlier elimination (Section 4.4).

After that key features are tracked (Section 4.5) in the subsequent images and used to estimate the camera motion. Two possible tracking methods could be chosen.

(a) To detect features in every stereo frame, then track them in the next frame. As two adjacent stereo pairs are always captured at two close positions, though feature tracking is practicable, vehicle motion estimation might not be very accurate as the reconstructed 3D positions for the same 3D object obtained at two instants do not change a lot.

(b) To select some image frames as reference, detect the key features in the reference frames, then track these features from the reference to the current frame and estimate the camera motion. Though the motion estimation problem sounds to be solved, feature tracking is a big challenge when two frames are captured at different positions with large distance.

Thus, we propose to make use of the advantages of these two methods: feature tracking is used between adjacent stereo pairs, while motion estimation is used between reference stereo pairs. The first stereo pair is selected as the initial reference. Features in the reference stereo pair are tracked across several frames till the number or distribution conditions of features cannot be satisfied. Then the stereo pair before current frame is selected as the new reference.

The previous steps will be repeated till the end of the sequence. With the reconstructed 3D points obtained by detected image features and tracked features, the relative camera motion between current and reference stereo pairs is estimated by a 3D transformation with RANSAC-based least square method (Sections 5.1, 5.2, and 5.3).

Finally, GPS positions are used to adjust the vision-based vehicle positions. It is necessary to check if they are accurate or fault by calculating the covariance matrices of GPS positions according to NMEA GST sentences (Section 5.4). If accurate GPS positions are available, they are used to adjust the camera positions. For the other part where accurate GPS signals cannot be received, only vision method is used. As the approach is based on stereovision, the considered stereoscopic system is briefly introduced in the next part.

4. 3D Landmarks Reconstruction and Tracking by Stereovision

Local 3D reconstruction and feature tracking based on rectified stereo images consists in four steps. At first, key features are separately extracted from the left and right images. Then, the correspondences are established under constraints. After that, the Z-depth and X, Y coordinates of the 3D points relative with current camera frame are estimated. Finally, key features in the reference stereo pair are tracked across several frames till the number or distribution conditions cannot be satisfied, the reference stereo pair is reinitialized, and the previous steps are repeated till the end of the sequence.

4.1. Stereoscopic System. A stereoscopic system is composed of two digital cameras C_1 and C_2. It permits to obtain two images of the same scene from two different perspectives simultaneously [28]. For the general stereoscopic systems, cameras are separately calibrated to obtain their intrinsic parameters, and also the relative position and orientation between the two cameras. Then the stereo images are rectified with known system parameters such that the general system can be transformed to an ideal stereovision system as in Figure 2, with camera focal length f, principal point (c_x, c_y) and baseline T connecting the optical centers C_1 and C_2. Given a 3D scene point $Q\{Q_x, Q_y, Q_z\}$ in the world coordinate system, it can be projected into two image points $q\{x, y\}$ and $q'\{x', y'\}$, respectively, on the left and right image planes, while q and q' are defined from their own image coordinate systems associated with the left and right cameras.

4.2. Features Extraction. In the 3D reconstruction literature, Harris corners [29] are often used as feature detector. However, they are sensitive to the changes of image scale, point of view, and illumination and cannot resist well to the noise. In order to obtain a good result of matching and tracking, the extracted features and descriptors must be robust and distinctive under various conditions (e.g., different light conditions, distances, or angles of view). Therefore in this article, the SURF features [30] are extracted from left and right images, then matched by the descriptors.

The SURF algorithm is composed of three steps. (1) Scale spaces are implemented as image pyramids. The layers are obtained by filtering the image with gradually bigger masks, taking into account the discrete nature of integral images and the specific structure of the filters. (2) Given a point $x = (x, y)$ in an image I, the determinant of approximated Hessian matrix $H(x, \delta)$ in x at scale δ is calculated; then, a nonmaximum suppression in a $3 * 3 * 3$ neighborhood is applied to localize interest points in the image over scales. (3) After that, the descriptors of SURF features are estimated based on pixel properties. It describes a distribution of Haar-wavelet responses within the interest point neighborhood.

4.3. Features Matching. Image feature matching is to measure the similarity between two features, then to find a list of corresponding feature couples between two images. In order to realize the matching, we can use several methods based on the nature of pixel information or descriptors.

The descriptor vectors can be matched based on a distance between the vectors (e.g., Mahalanobis distance, Euclidean distance). In this article, the SURF vectors with smallest Euclidean distance are considered as the corresponding features. Though the descriptors-based feature matching is very reliable, some false correspondences are still unavoidable. So after finding the closest corresponding features by SURF descriptors, ZNCC (Zero-mean normalized cross correlation) constraint and several geometric constraints are used to check the similarity of gray values between two features,

$$\mathrm{ZNCC} = \frac{\sum_{u,v\in W}\left(I_1(u,v)-\bar{I}_1\right)\left(I_2(u,v)-\bar{I}_2\right)}{\sqrt{\sum_{u,v\in W}\left(I_1(u,v)-\bar{I}_1\right)^2\sum_{u,v\in W}\left(I_2(u,v)-\bar{I}_2\right)^2}} \quad (1)$$

where W is the matching window centered around the image feature point; here, we choose the window size as $13 * 13$; $I_1(u, v)$ is the intensity of pixel within the window and $\bar{I}_1$ is the average intensity of the window in the left image; $I_2(u, v)$ is the intensity of pixel within the window and $\bar{I}_2$ is the average intensity of the window in the right image. For ZNCC, larger value indicates a closer relationship.

In order to improve the matching precision and computation time, some other geometric constraints are considered to reduce the search space and to refine the matching result.

(a) Epipolar constraint: for one image feature, its corresponding feature in another image must be within the relative epipolar line. For the rectified image pairs, the epipolar line is simplified as it is parallel to the horizontal scanlines.

(b) Maximum and minimum disparity constraint: according to the maximum and minimum depth of 3D objects, the disparity range can be estimated.

(c) Threshold of correlation score: only feature couples with the largest value of ZNCC > 0.9 will be chosen as the potential corresponding features.

(d) Uniqueness constraint: one feature can only be matched with another one.

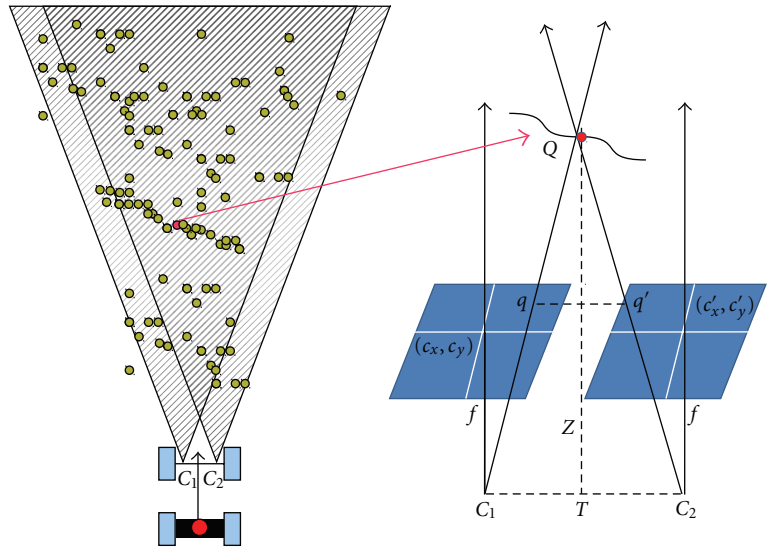


Figure 2: Ideal stereoscopic system.

(e) Ordering constraint: for corresponding couples $m \leftrightarrow m'$ and $n \leftrightarrow n'$, the position (left or right) of m with respect to n and the position of m' with respect to n' should be the same.

(f) Inverse matching (mutual checking): for every potential feature correspondence, taking the corresponding feature (x_{left} − disparity, y) in the right image as reference, search for its corresponding feature in the left image. If the same left image feature could be found, the matching result can be considered as an inlier.

4.4. Local 3D Reconstruction of Reference Stereo Pair

4.4.1. 3D Landmark Reconstruction of Reference Pair. When the geometric arrangement of the stereoscopic system is known, the local 3D position $Q(Q_x, Q_y, Q_z)$ of a point relative to the camera center can be recovered based on its corresponding image SURF features [28]. Typically, for a rectified stereoscopic system, the x-disparity information are directly used to reconstruct the 3D points. But due to the influence of various noises, the left and right rays passing through camera centers and corresponding features might not intersect at the same 3D point. The alternative method is to obtain the left and right rays separately based on the image features, find the shortest segment that connects these two rays, and take the middle point of this segment as the corresponding 3D position.

Let r_1 (resp., r_2) be the unit vector that connects the left (resp., right) camera center C_1 (resp., C_2) and corresponding left (resp., right) image feature q (resp., q'), Q_1 and Q_2 the endpoints of the shortest line segment connecting these two rays [1]. T is the baseline between C_1 and C_2, then

$$r_1 = \frac{\{q_x, q_y, f\}}{||C_1 q||}, \qquad r_2 = \frac{\{q'_x, q'_y, f'\}}{||C_2 q'||}, \tag{2}$$

and the relative distance between the 3D point and the two camera centers can be written as

$$Q_1 = C_1 + r_1 m_1, \qquad Q_2 = C_2 + r_2 m_2, \tag{3}$$

while $m_1 = \|Q_1 C_1\|$ and $m_2 = \|Q_2 C_2\|$, then

$$\begin{aligned} m_1 &= \frac{T \cdot r_1 - (T \cdot r_2)(r_1 \cdot r_2)}{1 - (r_1 \cdot r_2)^2}, \\ m_2 &= (r_1 \cdot r_2) m_1 - T \cdot r_2, \end{aligned} \tag{4}$$

then the coordinates of the 3D point Q can be obtained as

$$Q = \frac{(Q_1 + Q_2)}{2}. \tag{5}$$

We make a comparison of the reconstruction results by this middle-point method and the direct triangulation method with only x-disparity (III.B). In Figure 3, with the same feature correspondences, the Z-depth obtained by direct triangulation method is considered as the x-coordinates, and the corresponding Z-depth obtained by middle-point method illustrated above is considered as the y-coordinates. As in Figure 3(a), the Z-depth obtained by two methods seem to be the same. But when we zoom in, it demonstrates that when the point depth increases, the difference between the two methods grows: when the depth is less than 2 meters, the difference is all less than $4 * 10^{-3}$ m as in Figure 3(b); it increases to about 0.1 m when the depth increases to 8 meters as in Figures 3(c) and 3(d), when the depth increases to 20 m, the difference grows to about 0.6 m. Such reconstruction errors cannot be neglected. That is why we employed the middle-point triangulation method in this article though the computation is more complex.

After that, the distribution of image features in reference stereo pair is calculated as $\text{distribut}_{\text{reference}}$ by dividing the left image plane into a series of 24×24 squares and calculating the number of squares that contain the key features.

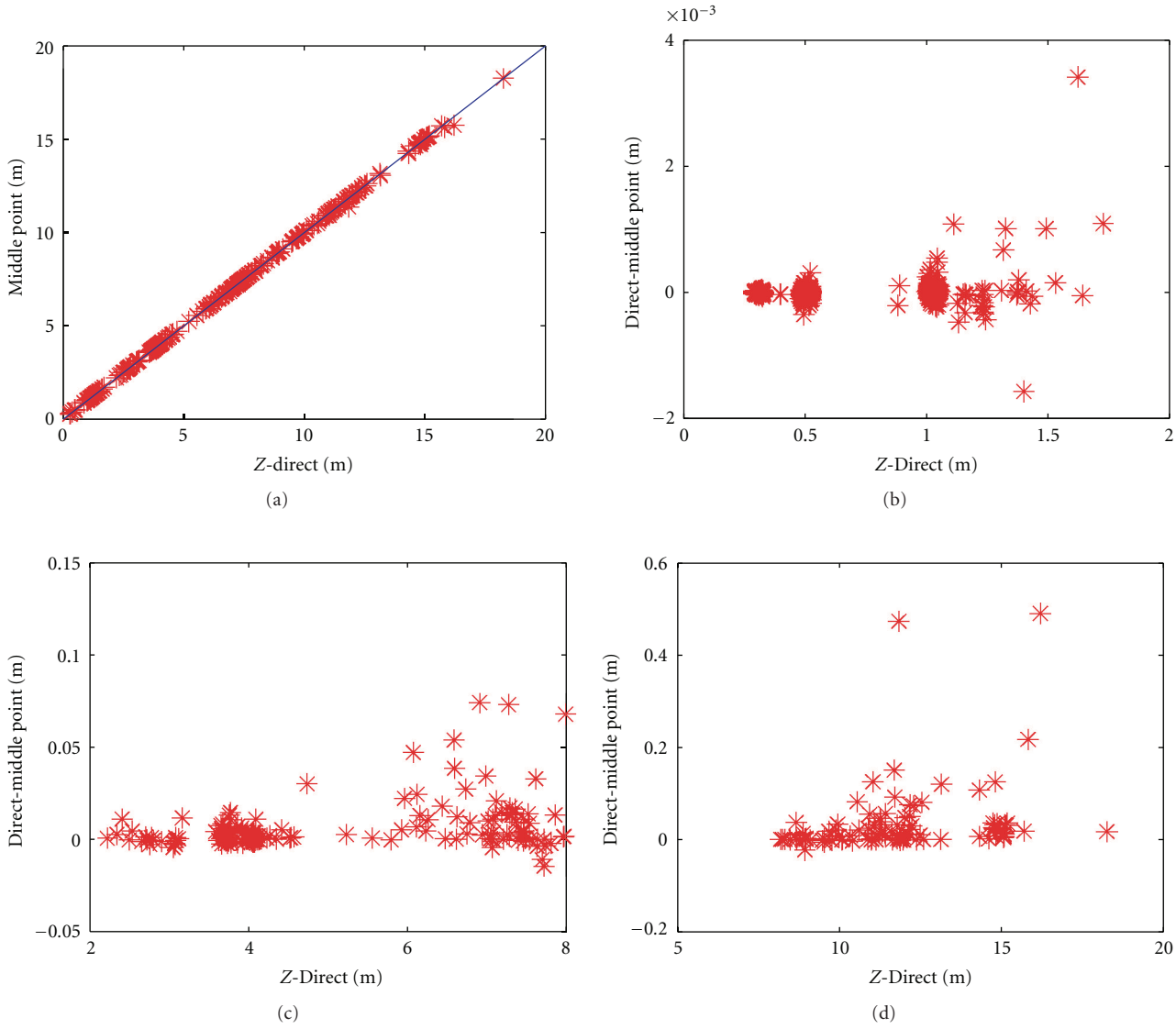


Figure 3: Difference between the direct and middle point triangulation methods.

4.4.2. Elimination of Landmarks with High Uncertainty. In order to achieve accurate navigation over long distances, errors in landmark matching and reconstruction processes must have a very small effect on every motion estimation step. Sources of 3D landmark outliers vary from the inaccurate camera calibration and physical image noise to the detection of features and matching errors, these will lead to some 3D points that do not really exist. When these landmarks are used, there is no doubt that the result would be ruined. Thus, we need to eliminate the 3D points that are less precise [31].

For a given stereo system with fixed intrinsic and extrinsic parameters (baseline), during the 3D reconstruction procedure, using the direct triangulation model $Z = f * T/d$, the derivative of Z to d is $\partial Z/\partial d = -fT/d^2$, replace d by $d = fT/Z$, the relationship of depth and disparity can be written as: $\partial Z/\partial d = Z^2/fT$, $\partial Z = (-Z^2/fT) \times \partial d$. Though the uncertainty induced by triangulation is not a simple scalar function of distance to the point as it is also skewed and oriented [32], we can note that the influence for the same uncertainty of disparity will become greater as the depth increases, and the accuracy of the depth estimation will decrease. Therefore, in our experiments, reconstructed 3D points with negative depth or with depth over 50 meters are eliminated. After eliminating the 3D outliers, 3D landmarks of reference stereo frames are stored. In the future work, we will take into account other factors for more accurate estimation.

4.5. Multiframes-Based Temporal Features Tracking

4.5.1. 2D Feature Tracking and Outlier Removing. In order to find the corresponding features in continuous frames, there are two alternative methods: the classic method is to detect features for every new frame, then use a matching method between frames. Another method is to use optical

flow by detecting the features in the first frame and then using a tracking method to track these features and to obtain their position in the current frame. As the geometric relationship between consecutive frames is unknown, both the computation of epipolar line and the use of ZNCC are method are time consuming; thus we choose Kanade-Lucas-Tomasi feature tracker [33] to track the key features.

2D features in the reference stereo pair are tracked across several frames until the number or distribution conditions of tracked features cannot be satisfied. The first stereo pair is selected as the initial reference stereo frame. When a new stereo pair is captured, the previous matched key features are separately tracked in the left and right images. During the tracking process, when false tracking occurs, there must be some mechanisms to detect and discard them. Four constraints are applied.

(a) Intensity constraint: the similarity SAD (Sum of Absolute Differences) of image features in the consecutive frames should be less than 500.

(b) Search space constraint: the tracked feature cannot move out of the current image plane, $0 < x <$ image.x, $0 < y <$ image.y.

(c) Epipolar constraint of tracked features in the current left and right frames.

(d) ZNCC threshold of tracked image points in the left and right images (the same constraints as the Feature Matching part).

Furthermore, a new relative depth constraint is added. When the vehicle moves in rigid and static environment, the changes of estimated depths relative to the camera frames t and $t+1$ should be almost the same for all the reconstructed 3D points according to

$$\Delta Z = Z(t) - Z(t+1) = f * T * \left(\frac{1}{d(t)} - \frac{1}{d(t+1)}\right). \quad (6)$$

Therefore, the mean and standard deviation of relative depth changes are calculated for all the tracked couples, and the couples whose depth deviation is more than 3 times of the standard deviation are discarded. This method can be used to eliminate some dynamic objects in the environment with a different velocity, for example, pedestrians, other moving vehicles, and so forth.

After obtaining the tracked features in the left and right frames, the distribution of features in the current stereo frame is calculated as $\text{distribut}_{\text{tracking}}$. Then, the 3D position of tracked image features in current camera coordinate system is estimated.

4.5.2. Reference Stereo Pair Selection and Updating. As the camera moves, some features may move out of the field of view and some will be rejected as outliers; only features that can be tracked by the previous frame will be tracked sequentially. Thus, this feature number will decrease across the frames. The proposed reference stereo pair selection and updating mechanism is illustrated in Figure 4. For accurate pose estimation, enough features number and spatial distribution should be ensured. If the matched points number of the reference stereo pair is d, the number threshold of tracked points is set as $\text{Threshold}_{\text{number}} = d * 60\%$; when sometimes the number d is too small, this threshold is set to 10. Then, we check the spatial distribution of tracked feature pairs with a threshold as $\text{Threshold}_{\text{distribut}} = (\text{distribut}_{\text{reference}} * 60\%)$. If the number or distribution is less than the threshold, the previous stereo pair is selected as new reference stereo pair, then SURF features are detected and the previous procedures are repeated. If two frames are consecutive, but the tracked number is less than the threshold, we still use the previous frame as reference.

At the same time, the detected features in the new reference stereo pair are compared with those of the previous reference pair, and the new appeared landmarks are added into the 3D model.

5. Vehicle Egomotion Estimation

5.1. Vision-Based Relative Motion Estimation. Assume that the ground is plane, the camera motion can be represented by X-Z coordinates and yaw angle θ. For two corresponding point sets $\{Q_i^{t+1}\}$ and $\{Q_i^{\text{ref}}\}$ ($i = 1 : N$, while N is the number of corresponding points) obtained at the current camera coordinate system $t+1$ and the reference coordinate system, the two point sets can be related by

$$\left\{Q_i^{t+1}\right\} = R * \left\{Q_i^{\text{ref}}\right\} + T + V_i, \quad (7)$$

where R is a standard $2 * 2$ rotation matrix $\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$, $T(T_x, T_z)$ is a 2D translation vector, and V_i is the noise vector. To find the optimal transformation $[R, T]$ that transforms the previous set onto current set, it requires to minimize the residual error

$$\varepsilon^2 = \sum_{i=1}^{N} \left\| Q_i^{t+1} - RQ_i^{\text{ref}} - T \right\|. \quad (8)$$

With the assumption that the environment is rigid, the centroid of two point sets should be the same

$$\begin{aligned} \overline{Q}^{t+1} &= \frac{1}{N}\sum_{i=1}^{N} Q_i^{t+1}, \qquad \left\{Q_{ci}^{t+1}\right\} = \left\{Q_i^{t+1}\right\} - \overline{Q}^{t+1}, \\ \overline{Q}^{\text{ref}} &= \frac{1}{N}\sum_{i=1}^{N} Q_i^{\text{ref}}, \qquad \left\{Q_{ci}^{\text{ref}}\right\} = \left\{Q_i^{\text{ref}}\right\} - \overline{Q}_i^{\text{ref}}, \\ \varepsilon^2 &= \sum_{i=1}^{N} \left\| Q_{ci}^{t+1} - R \cdot Q_{ci}^{\text{ref}} \right\|^2 \\ &= \sum_{i=1}^{N} \left({Q_{ci}^{t+1}}^T Q_{ci}^{t+1} + {Q_{ci}^{\text{ref}}}^T Q_{ci}^{\text{ref}} - 2{Q_{ci}^{t+1}}^T R Q_{ci}^{\text{ref}} \right). \end{aligned} \quad (9)$$

This equation is minimized when the last term is maximized. This is equivalent to maximizing the trace (RH), the cost expression is written as

$$H = \sum_{i=1}^{N} \left(Q_{ci}^{\text{ref}} {Q_{ci}^{t+1}}^T \right). \quad (10)$$

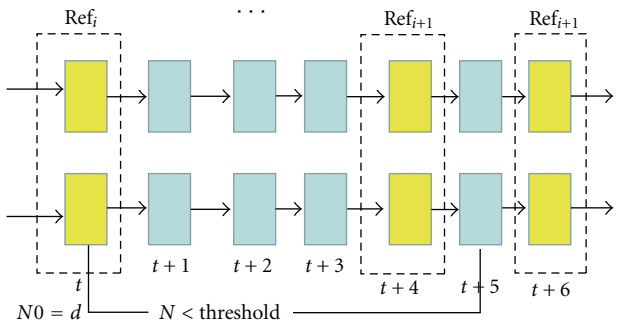


Figure 4: Selection and reinitialization of reference image pair.

The best rotation in the least squares sense can be found by SVD (singular value decomposition) of matrix $H = UDV^T$ as Arun et al.'s solution [34] together with Umeyama's complement [35] for some degenerated cases:

$$S = \begin{cases} I, & \text{if } \det(V) * \det(U) = 1, \\ \text{diag}(1, -1), & \text{if } \det(V) * \det(U) = -1. \end{cases} \tag{11}$$

Then, rotation matrix can be obtained by $R = VSU^T$. The translation vector T is: $T = \overline{Q}^{t+1} - R \cdot \overline{Q}^{\text{ref}}$, as the optimal translation vector aligns the centroid of the set $\{Q_i^{t+1}\}$ by the rotated centroid of the set $\{Q_i^{\text{ref}}\}$.

5.2. Iterative Motion Estimation by RANSAC. Some 3D points with large depth uncertainty are already rejected before motion estimation, then RANSAC [36] is used for a more precise estimation of the model parameters. It achieves the goal by iteratively selecting a random subset of three tracked image points in current frame (left), triangulating the points, and then generating one camera motion hypothesis R and T. Inliers are determined by transforming the reconstructed points in reference coordinate system into current camera system based on the new motion hypothesis and then calculating the Euclidean difference between the two 3D positions, the best hypothesis is the one with the largest number of inliers. And after that, the final solution can be obtained by using the largest inlier subset.

To ensure that the randomly selected three points distribute well in the image, every two image features must have a distance large than the square size 24. The other parameters are dynamically chosen according to [37]: for a probability of 95% that a point (correspondence) is an inlier, the distance threshold to determine whether a point correspondence is an inlier is set as $6 * \delta^2$, where δ is the standard deviation of measurements; the number of sampling times N_{sample} is dynamically estimated by $N_{\text{sample}} = \log(1-p)/\log(1-(1-\epsilon)^s)$ to ensure a probability p that at least one of the random samples of 3 points is free from outliers, while p is set as 0.99, and ϵ is the percentage of outliers, where $\epsilon = 1 - (\text{number}_{\text{inliers}})/(\text{number}_{\text{points}})$.

5.3. Global Camera Motion Based on Incremental Procedures. Taking the first camera position as the origin of global coordinate system, the global motion of every camera position and the global 3D landmarks can be obtained based on the reference stereo pair

$$\begin{aligned} R_{\text{global}} &= R_{\text{ref}} * R_n, \\ T_{\text{global}} &= R_{\text{ref}} * T_n + T_{\text{ref}}, \\ \left\{Q_n^{\text{ref}}\right\} &= R_{\text{global}} * \{Q_n^t\} + T_{\text{global}}, \\ \left\{Q_n^{\text{global}}\right\} &= {R_{\text{global}}}^T * \left(\left\{Q_n^{\text{ref}}\right\} + T_{\text{global}}\right). \end{aligned} \tag{12}$$

With R_{ref}, T_{ref}: rotation matrix and translation vector of reference stereo pair in global system; R_n, T_n: camera rotation matrix and translation vector of frame t relative with its reference stereo pair; R_{global}, T_{global}: camera rotation matrix and translation vector in global system; $\{Q_n^t\}$: 3D points relative with local stereo pair; $\{Q_n^{\text{ref}}\}$: 3D points relative with the current reference of stereo pair (left camera center); $\{Q_n^{\text{global}}\}$: 3D points in the world coordinate system.

5.4. Vehicle Trajectory Estimation by Integration of GPS and Vision Data

5.4.1. Uncertainty of GPS Position. In the urban environment, GPS suffers from multipath problems, and the nonstationary of the GPS measurement noise affects the observation model. As GPS points are in a known coordinate system, the linear observation equation is in the form

$$P_k = \begin{bmatrix} x_k^{\text{gps}} \\ z_k^{\text{gps}} \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_k \\ Z_k \\ \theta_k \end{pmatrix} + \beta_{\text{gps}}, \tag{13}$$

while the GPS observation $(x_{\text{gps}}, z_{\text{gps}})$ is provided by GPS position measurement and β_{gps} is the measurement noise.

As GPS measurements are affected by many independent noise sources, assume that the position vector $P = [P_1, P_2, \ldots, P_n]$ has a Gaussian distribution, the covariance matrix Q_k^{gps} of every GPS position error can be estimated by

$$Q_k^{\text{gps}} = \begin{pmatrix} \delta_{x,\text{gps}}^2 & \rho \cdot \delta_x \cdot \delta_z \\ \rho \cdot \delta_x \cdot \delta_z & \delta_{z,\text{gps}}^2 \end{pmatrix}. \tag{14}$$

While $\delta_{x,\text{gps}}$ and $\delta_{z,\text{gps}}$ are the standard deviations of the estimation error of x and z as observed in the x-z plane, ρ is the spatial correlation coefficient, and φ is the orientation of semimajor axis of error ellipse in degrees from true North. $\delta_{x,\text{gps}}$, $\delta_{z,\text{gps}}$, and φ can be obtained by the Standard National Marine Electronics Association (NMEA) sentence "GST", and ρ can be calculated by

$$\rho = \begin{cases} \dfrac{tg(2\varphi)(\varphi_x^2 - \varphi_z^2)}{2\varphi_x\varphi_z}, & 0 < \varphi < \dfrac{\pi}{2}, \\ \dfrac{tg(2\varphi)(\varphi_z^2 - \varphi_x^2)}{2\varphi_x\varphi_z}, & -\dfrac{\pi}{2} < \varphi < 0. \end{cases} \tag{15}$$

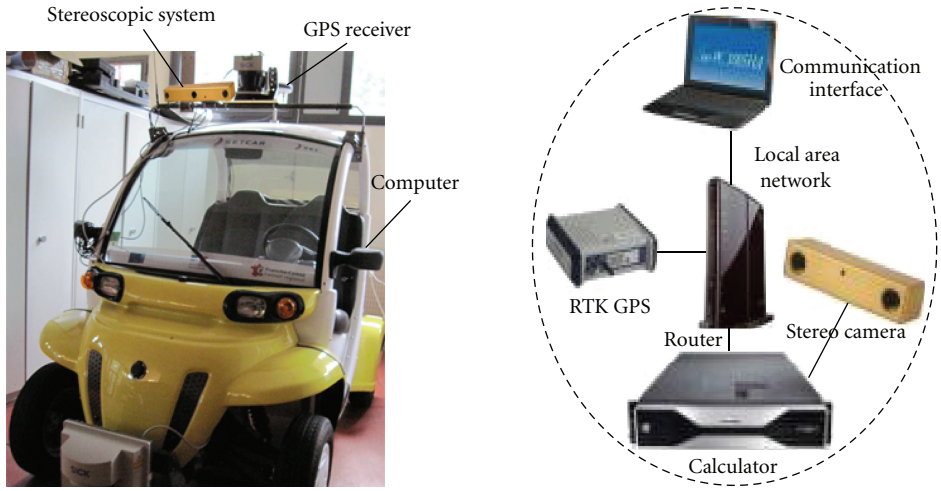


Figure 5: (a) Experimental vehicle equipped with RTK-GPS and stereoscopic system. (b) Hardware architecture.

5.4.2. Integration of GPS and Vision Position. Two sensor systems should be under the same timing system and coordinate system. As mentioned before, the initial vehicle position is considered as the origin of the global system. The GPS trajectory is translated into the vision system by transforming the corresponding initial GPS position to $\{0,0\}$. Then, based on the first 30 positions, the vision-based trajectory is rotated such that its forward direction is the same as the GPS trajectory,

$$Q_{\{0,0\}}^{\text{global}} = P_{\{0,0\}}^{\text{global}} + T_{\text{gps}}^{\text{vision}},$$
$$Q_{\text{vision}}^{\text{gps}} = R_{\text{vision}}^{\text{gps}} * Q^{\text{global}}. \quad (16)$$

With $Q_{\{0,0\}}^{\text{global}}$: initial vehicle position as origin $\{0,0\}$; $P_{\{0,0\}}^{\text{global}}$: global position of the corresponding initial GPS position; $T_{\text{gps}}^{\text{vision}}$: translation vector from initial GPS position to the origin of vision coordinate system $\{0,0\}$; Q^{global}: global position of vision-based trajectory; $R_{\text{vision}}^{\text{gps}}$: rotation matrix from vision-based trajectory to GPS trajectory; $Q_{\text{vision}}^{\text{gps}}$: global position of vision-based trajectory in the same forward direction as GPS.

Suppose that during some periods the GPS data cannot be received, only the vision based method is used to estimate the camera motion. At the same time, the precision of GPS position is checked once every 50 GPS positions (or the other regular interval) according to the NMEA GPGST sentence by calculating the covariance matrix of GPS position. If it is accurate, GPS data is used to adjust the trajectory obtained by cameras, including the vehicle orientation and the position. In order to evaluate the performance of the proposed localization method, two video sequences with ground-truth were tested.

6. Experimental Results

The proposed approach was implemented in C and C++, with the Open Source Computer Vision library OpenCV and the public domain linear algebra package LAPACK. As shown in Figure 5, our experimental GEM vehicle is equipped with a stereoscopic Bumblebee XB3 camera (5 Hz) and a Magellan ProFlex500 RTK-GPS (10 Hz) receiver. They are mounted on the roof of the vehicle.

The RTK-GPS receiver can achieve up to 1 cm accuracy in an horizontal plane. Cameras are calibrated and images are rectified. The Bumblebee XB3 has three collinear cameras, with 0.12 m distance apart from each other. Every image has a resolution of $1280 * 960$. The left and right cameras are used as a stereoscopic system with baseline 0.24 m. GPS and image data are stored under the same computer together with their stored times. Then, the synchronization of the two sensor systems are achieved by associating their saving times.

6.1. GPS Transformation. As GPS provides longitude and latitude information in Earth frame, the GPS positions obtained from NMEA GPGLL sentences are projected from WGS84 system to Extended Lambert II coordinate system that covers the region of Belfort (http://professionnels.ign.fr/ficheProduitCMS.do?idDoc=5352513). Then, they are translated into the vision frame by transforming the corresponding initial GPS position to $\{0,0\}$. GPS positions could then be shown in an XY geographical coordinate system and compared with the vision-based results. After that, based on the initial vehicle orientation obtained by VO and GPS, the vision-based trajectory is rotated such that its forward direction is the same as the GPS trajectory. The transformed RTK-GPS trajectory will be served as the ground-truth to evaluate the performance of the proposed visual odometry method and RTK-GPS/VO integrated method.

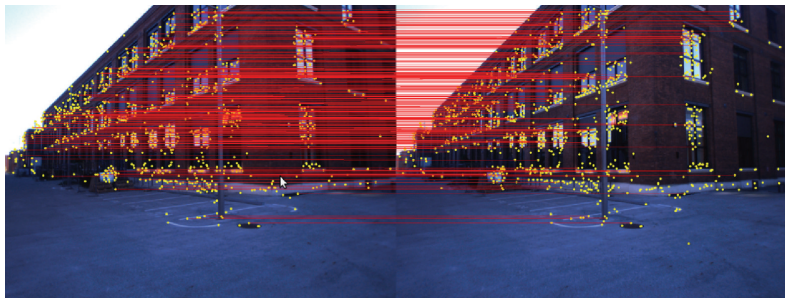
6.2. Feature Detection and Landmarks Reconstruction Results. In this part, we illustrate the results of 3D mapping as explained in Section 4. At first a new pair of images are captured by the calibrated stereoscopic system on-board; then, the SURF features are extracted from the two images;

(a) Left image from the sequence

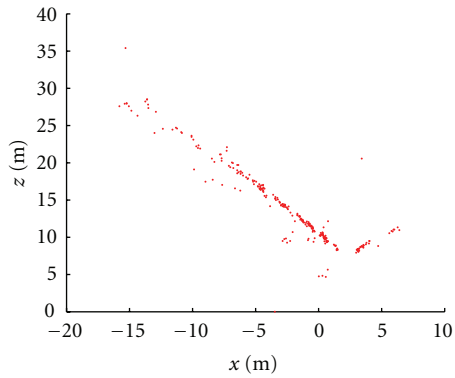
(b) The corresponding right image

(c) The detected SURF features in left image

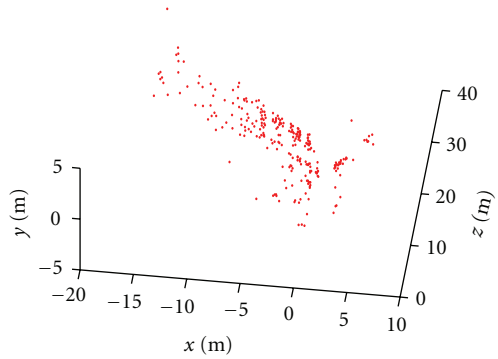
(d) The detected SURF features in right image



(e) The 2D/2D matching results of SURF feature in left and right images are connected with red lines



(f) Reconstructed 3D points, overlook



(g) Reconstructed 3D points, side view

Figure 6: The yellow points are the detected features; the corresponding features are connected by red line; number of detected SURF features in left image is 1048, in right image is 1142; the stereo matched points number of image pair is 334, and distribution score is 225.

Figure 7: The bottom image pair is the reference images pair, yellow points indicate the detected features, and red features are the ones that could be tracked; the above image pair is the tracked image pair, the yellow points are the positions of tracked features in current frame, and the change of feature position is demonstrated by red line. The number of tracked points in left image is 332, the initial number of tracked points in right image is 320, number after 2D filtering is 299, and distribution score for current left frame is 207.

after that, the SURF features are matched by the descriptors. Then, ZNCC and some geometric constraints are used to eliminate the bad matching results.

We select one image pair from the real sequence and show the results of every step of the reconstruction method in Figure 6. Extracted features number in left image is 1048, in right image is 1142, and the stereo matched points number is 334, distribution score is 225, so the new 3D points are 334. Such quantity of points is sufficient for the 3D reconstruction. They are reconstructed into 3D space as landmarks.

Then, these matched features are tracked in the next image pair. The tracking result is shown in Figure 7. The bottom image pair is the reference images pair, yellow points indicate the detected features, and red features are the ones that could be tracked; the above image pair is the tracked image pair, the yellow points are the positions of tracked features in current frame, and the changes of feature positions are demonstrated by red lines. The number of tracked features in left image is 332, the number of initial tracked features in right image is 320, and after the 2D outlier elimination, the number of remaining tracked features is 299, so the number of tracked feature couples is 299, and the distribution score for current left frame is 207. Then, these features are also reconstructed, and used to estimate vehicle motion.

6.3. Localization Result: First Experiment. The first sequence was captured in May, 2010, at Belfort, France. In Figure 8, the RTK-GPS trajectory is presented with blue line on Google Earth. The total length of GPS trajectory is about 290 m.

6.3.1. Stereovision Method-Based Results. At first, we estimated the vehicle trajectory only with stereovision method. The vision-based trajectory was transformed to the same coordinate system as RTK-GPS. As illustrated in Figure 13(a), the blue line and red line indicate the vehicle trajectories separately obtained by RTK-GPS and stereovision method. The trajectory obtained by stereovision is 251.4 m, with 13.43% error of the total length. The vision-only method works well at short term, but as the errors accumulate gradually, the trajectory drifts. As every GPS position was associated with a vision position by their closest saving time, the differences of vehicle positions in both X and Z directions are shown in Figure 13(c). The errors increase progressively, with mean error as 25.87 m and standard deviation as 17.05 m. We also compute the yaw angle of the vehicle at every instant. As no IMU sensor was incorporated in our system, the ground truth of yaw angles was approximated by calculating the angles between two subsequent GPS positions though they are very rough.

Figure 8: The first RTK-GPS trajectory (on satellite image of Google Earth).

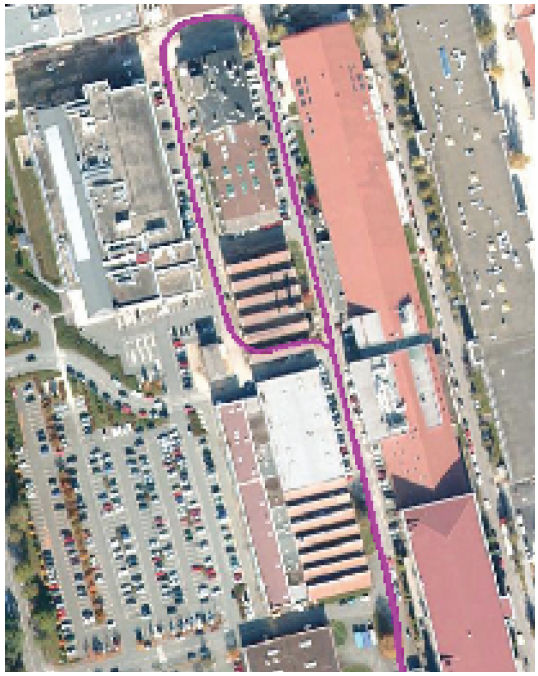


Figure 9: The second RTK-GPS trajectory (on satellite image of Google Earth).

The difference of orientation is shown in Figure 13(e), with mean error as 7.27° and standard deviation as 14.10°.

6.3.2. GPS/Vision Integrated Method. Then, we tested the same sequence with integration method of GPS and stereovision. Based on the NMEA GPGLL sentences of GPS, GPS positions could be obtained at every instants. As no GPGST sentences are stored for this sequence, we adjusted the vision-based vehicle trajectory once every 50 GPS positions: both of the GPS orientation and position were given to the vision point. The GPS and adjusted vision trajectory are shown in Figure 13(b) with blue and red lines, respectively.

It illustrates that the vision trajectory (red) can fit the GPS trajectory (blue) better than the stereovision-only method, especially for the orientation. As in Table 1, the GPS/Vision integration method-based trajectory is 294.48 m, so the difference is 1.41%. The difference in X and Z direction are compared Figure 13(d). We can see that most of the position differences in x and z direction are both less than 2 m, with mean error of position as 1.12 m and standard deviation as 0.87 m. The mean error of orientation difference is 1.70° and standard deviation as 12.71°. The result is more precise than the vision-only method.

Table 1: Comparison of traveling distance (/m).

Method	Ground truth	Estimated dist.	Error %	Mean	Std.
VO	290.40	251.40	13.43	25.87	17.05
GPS/VO	290.40	294.48	1.41	1.12	0.87

6.4. Localization Result: Second Experiment. Another sequence was captured in September, 2010, at Belfort, France. The vehicle was driven in an industrial area where there are buildings around as shown in Figure 9. The sequence is longer than the first one. It comprises 1800 stereo pairs. The RTK-GPS trajectory is presented with magenta line on *Google Earth*. Trajectory distance measured by RTK-GPS was about 651.78 m, with direct lines and four big turns with about 90°.

6.4.1. Stereovision Method-Based Results. In this section, we compared the trajectory obtained by 3D visual odometry (with Tx, Ty, Tz and yaw, pitch and roll angles) and 2D visual odometry (with only Tx, Tz and yaw). The 3D trajectory was projected onto X-Z plane and compared with the 2D trajectory obtained under ground plane assumption. Suppose that the ground is flat, the comparison of traveling distances for two methods can be seen in Table 2. As shown in Figure 10(b), both the trajectories obtained by 2D and 3D visual odometry drift gradually. Though the 3D odometry seems better than the 2D method on X-Z plane, the final absolute altitude by 3D visual odometry is unrealistic. For a travel about 700 m, the final altitude was about −30 m as shown in Figure 10(a). Thus, in the following part, 2D visual odometry and GPS are integrated to obtain a more accurate estimation on X-Z plane. The problem of accumulated error of altitude will be studied later by the combination of INS sensor for more precise angle measurements.

As mentioned in the previous part, the vehicle orientation estimated by the GPS position is not very accurate as the GPS signals jump sometimes. For this sequence, the NMEA GPVTG sentences are also stored. The parameter COG (orientation with respect to the True North) can be used to calculate the yaw angle of vehicle. In Figure 11, we compare the orientation directly obtained by GPS positions and GPVTG sentences. It shows that some large jumps exist for the GPS position-based method, so we use the COG orientation as ground truth. The difference of stereovision-based vehicle orientation and the ground truth is shown in Figure 14(e). With mean error as 10.59° and standard deviation as 6.45°.

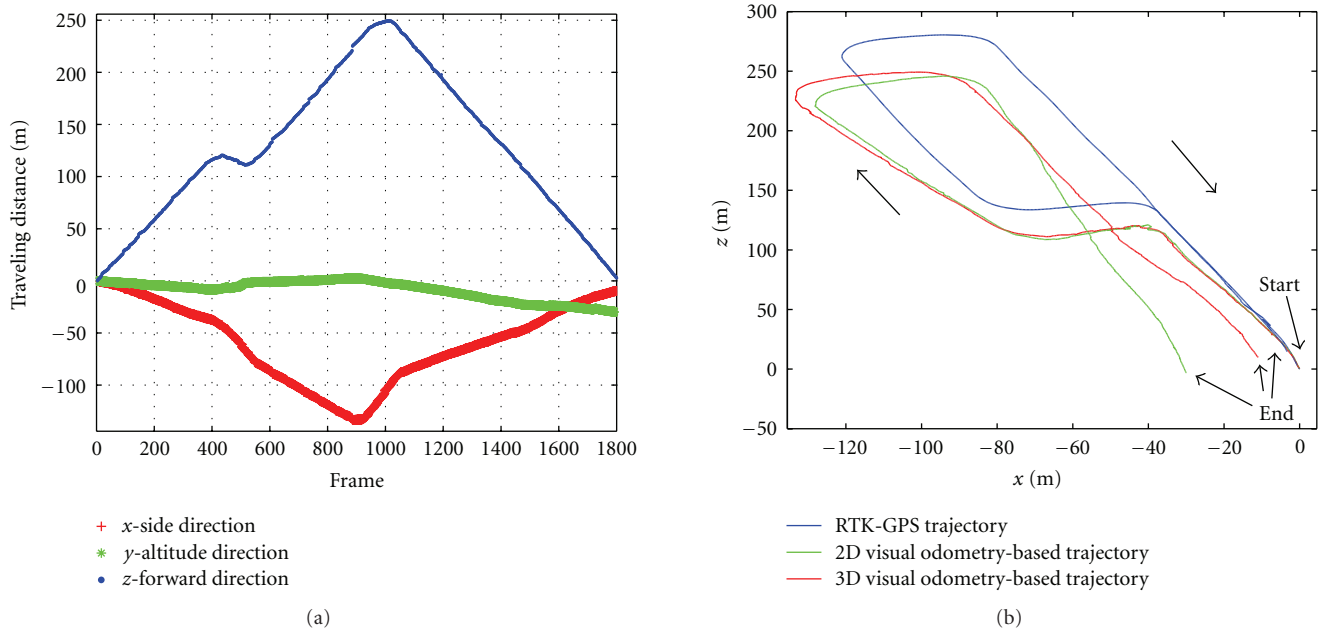


Figure 10: (a) Estimated vehicle motion by 3D visual odometry. (b) Trajectory obtained by 3D information and under 2D assumption.

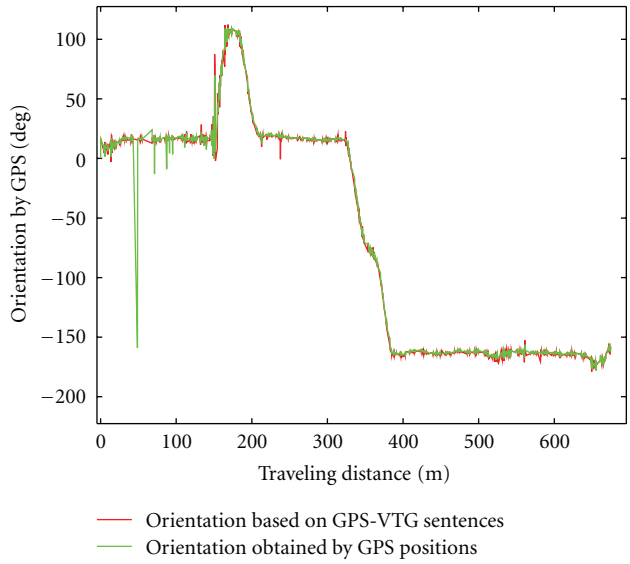


Figure 11: The vehicle orientation obtained by GPS positions and NMEA VTG sentences.

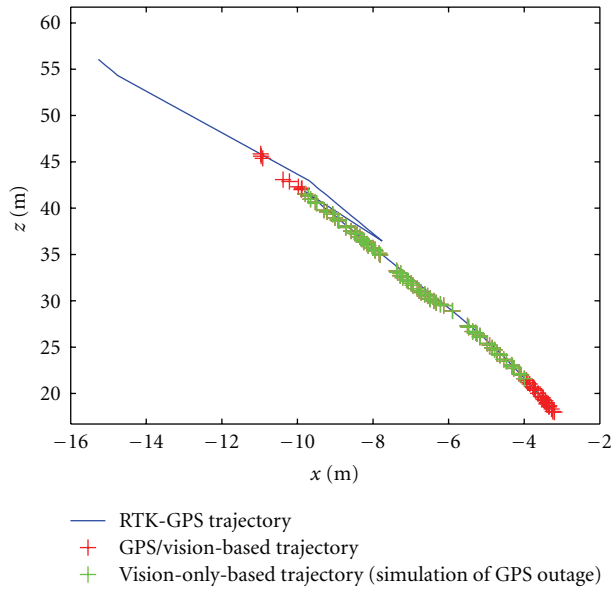
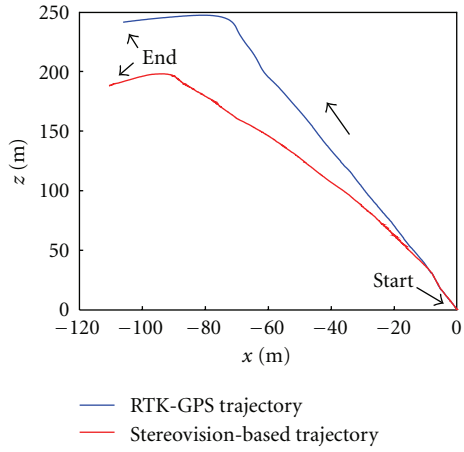


Figure 12: Estimated vehicle trajectory with and without GPS mask.

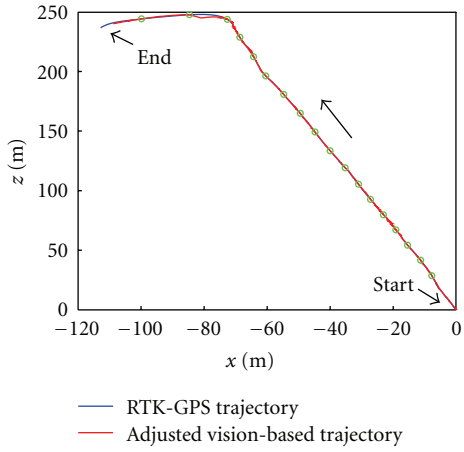
Table 2: Comparison of traveling distance (/m).

Method	Ground truth	Estimated dist.	Error %	Mean	Std.
Vision-2D	651.78	627.05	3.79	27.61	11.83
Vision-3D	651.78	636.46	2.35	25.06	10.11
GPS/VO	651.78	683.23	4.83	0.18	0.92

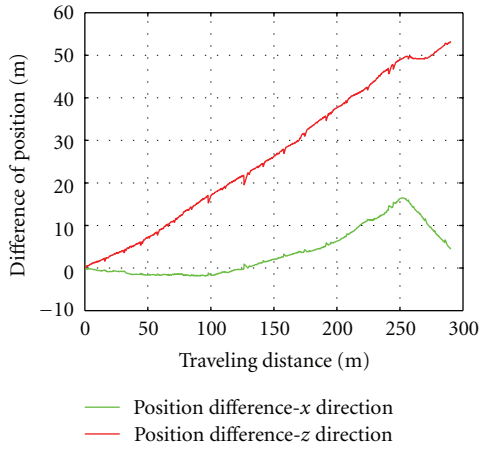
6.4.2. GPS/Vision Integrated Method. Then, we tested the same sequence with the integration method of GPS and stereovision. The first step is to check that whether the GPS signals can be received or not. If they are received, the precision of every GPS position is checked by covariance matrices based on NMEA GPGST sentences. If the first parameter of covariance matrix is less than 10 meters, this GPS position is used to adjust the stereovision-based trajectory. Results are shown in Figure 14(b). The total distance is 683.23 m. Though the difference of total length with ground truth is 4.83%, it should be noticed that the



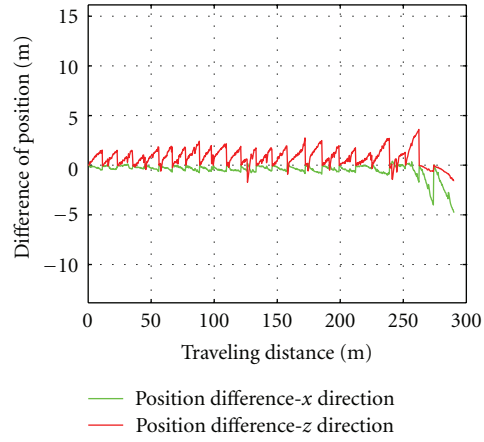
(a) Trajectories separately obtained by GPS and stereovision



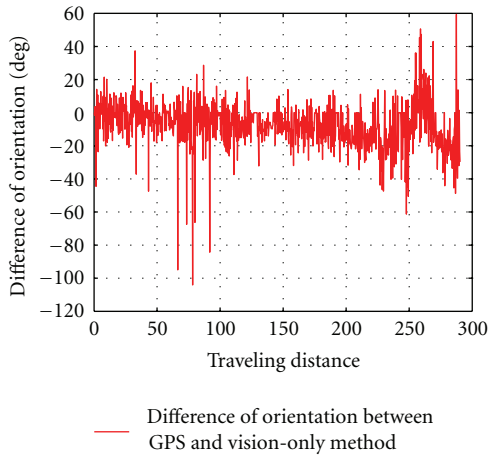
(b) Trajectory obtained by GPS/Vision integrated method



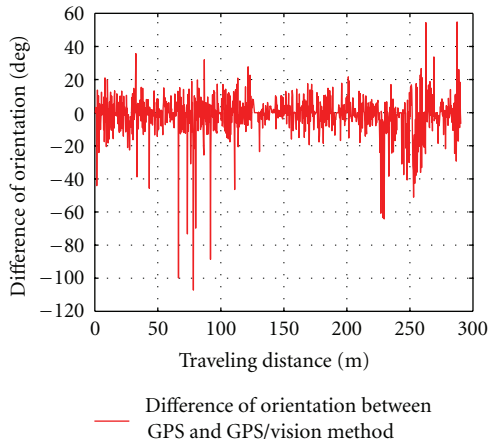
(c) Difference of position between GPS and vision-only trajectories



(d) Difference of position between GPS and GPS/Vision trajectories

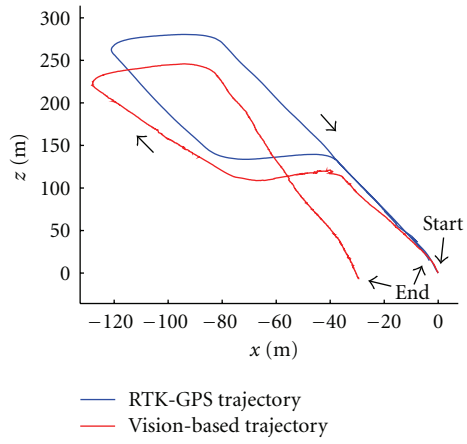


(e) Difference of orientation between GPS and vision-only trajectories

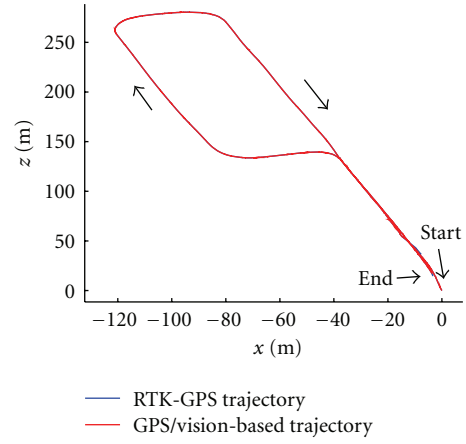


(f) Difference of orientation between GPS and GPS/Vision trajectories

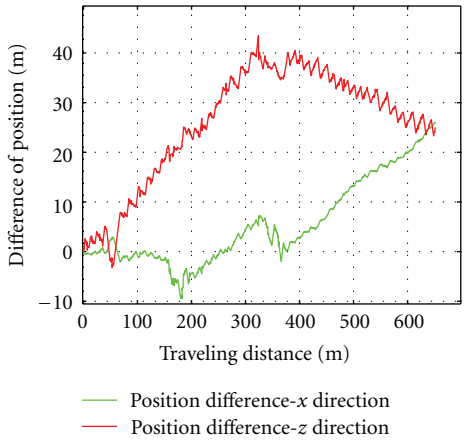
Figure 13: The first image sequence. Trajectories obtained by GPS, stereovision and GPS/Vision integrated methods. Difference of position and orientation between RTK-GPS and GPS/Vision trajectories.



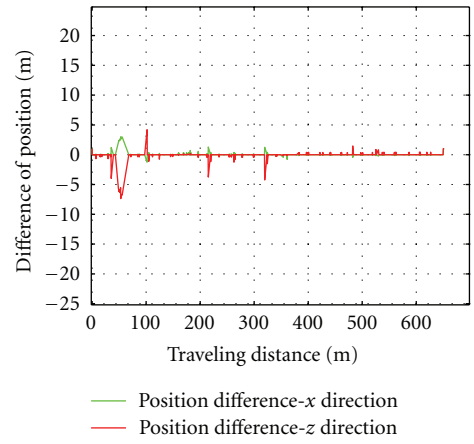
(a) Trajectories separately obtained by GPS and stereovision



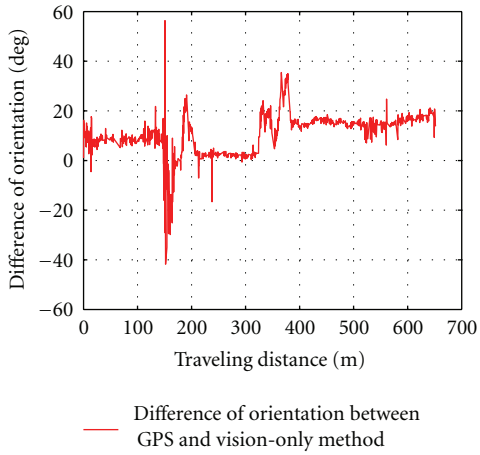
(b) Trajectory obtained by GPS/Vision integrated method



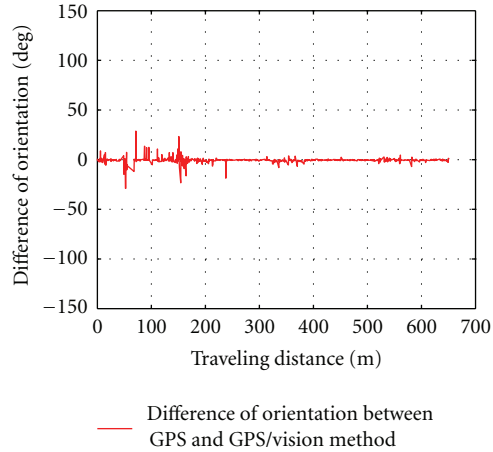
(c) Difference of position between GPS and vision-only trajectories



(d) Difference of position between GPS and GPS/Vision trajectories



(e) Difference of orientation between GPS and vision-only trajectories



(f) Difference of orientation between GPS and GPS/Vision trajectories

Figure 14: The second image sequence. Trajectories obtained by GPS, stereovision, and GPS/Vision integrated methods. Difference of position and orientation between RTK-GPS and stereovision-only trajectories.

mean error and standard deviation of position error are largely reduced. The mean error of orientation difference is only 0.18°, and the standard deviation is 0.92°.

6.4.3. Simulation of GPS Mask. A GPS mask about 50 measurements (the distance and time corresponding to 50 GPS measurements were 26.5 m and 7 s) was simulated in this section. No GPS signals were provided during the GPS masked areas, and only vision-based method was used. The result shows that during the process whereby there are GPS outages, the vision-based method can continue to estimate the vehicle positions. As shown in Figure 12, the red crosses indicate the vehicle positions when GPS measurements are available, and the green crosses indicate the vehicle positions when GPS signals are masked. It illustrates that the stereovision method can well estimate the vehicle motion at every instant. Thus, when accurate GPS position is not available (e.g., GPS jumps), the stereovision method can provide precise estimation in short term.

7. Conclusions and Future Works

7.1. Conclusions. We present a stereovision and GPS integrated method for 3D scenery mapping and vehicle localization in dense urban environments. This method is based on feature detection, matching and tracking, and construction of 3D environment and vehicle motion by integration of accurate GPS positions and stereovision. The integration of GPS can reduce the error accumulation of stereovision based visual odometry, especially for the large yaw errors in areas with less distinctive features (high way or a lot of repeating textures such as trees, etc.). And vice versa, the stereovision-based motion estimation can correct the inaccurate GPS measurements, for example, due to the multipath problem.

7.2. Future Works. As future works, we plan to incorporate other sensors into our system, such as using IMU to provide orientation and velocity, and using scanning laser range finders to perform mapping in large-scale environment. In this paper, during the test, the buildings and other objects are not far away from the vehicle, and enough features could be detected and tracked during the sequence. But if the vehicle quickly changes the orientation, or when the scenery objects in FOV are far away, or when the illumination is too strong or shadow exists, or on cloudy days, the detected features and tracked features might not be enough. If the tracked number is less than 3, the only use of stereovision-based egomotion method cannot work, thus in our future works, GPS position will not only be used to adjust the vision-based vehicle trajectory, but also be integrated to provide the vehicle position when vision becomes invalid using interpolation or motion model. Besides, in this paper, we assume that the ground is plane without considering the altitude, we will try to solve the 6DOF camera motion problem.

References

[1] Y. Cheng, M. W. Maimone, and L. Matthies, "Visual odometry on the Mars Exploration Rovers," *IEEE Robotics and Automation Magazine*, vol. 13, no. 2, pp. 54–62, 2006.

[2] Z. Chen and S. T. Birchfield, "Qualitative vision-based mobile robot navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '06)*, pp. 2686–2692, May 2006.

[3] E. Royer, M. Lhuillier, M. Dhome, and J. M. Lavest, "Monocular vision for mobile robot localization and autonomous navigation," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 237–260, 2007.

[4] P. Lothe, S. Bourgeois, F. Dekeyser, E. Royer, and M. Dhome, "Towards geographical referencing of monocular SLAM reconstruction using 3D city models: application to real-time accurate vision-based localization," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2882–2889, June 2009.

[5] S. Se and P. Jasiobedzki, "Stereo-vision based 3D modeling for unmanned ground vehicles," in *Unmanned Systems Technology*, vol. 6561 of *Proceedings of SPIE*, Orlando, Fla, USA, 2007.

[6] S. Nogueira, Y. Ruichek, and F. Charpillet, "A Self Navigation Technique using Stereovision Analysis," in *Stereo Vision*, pp. 287–298, InTech Education and Publishing, 2008.

[7] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM)—part II," *IEEE Robotics and Automation Magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[8] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, The MIT Press, 2005.

[9] B. Williams and I. Reid, "On combining visual SLAM and visual odometry," in *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3494–3500, 2010.

[10] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, pp. I652–I659, July 2004.

[11] D. W. Eggert, A. Lorusso, and R. B. Fisher, "Estimating 3-D rigid body transformations: a comparison of four major algorithms," *Machine Vision and Applications*, vol. 9, no. 5-6, pp. 272–290, 1997.

[12] T. Gandhi and M. Trivedi, "Parametric ego-motion estimation for vehicle surround analysis using an omnidirectional camera," *Machine Vision and Applications*, vol. 16, no. 2, pp. 85–95, 2005.

[13] J. P. Tardif, Y. Pavlidis, and K. Daniilidis, "Monocular visual odometry in urban environments using an omnidirectional camera," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '08)*, pp. 2531–2538, September 2008.

[14] R. García-García, M. A. Sotelo, I. Parra, D. Fernández, J. E. Naranjo, and M. Gavilán, "3D visual odometry for road vehicles," *Journal of Intelligent and Robotic Systems*, vol. 51, no. 1, pp. 113–134, 2008.

[15] C. F. Olson, L. H. Matthies, M. Schoppers, and M. W. Maimone, "Rover navigation using stereo ego-motion," *Robotics and Autonomous Systems*, vol. 43, no. 4, pp. 215–229, 2003.

[16] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry for ground vehicle applications," *Journal of Field Robotics*, vol. 23, no. 1, pp. 3–20, 2006.

[17] K. Konolige, M. Agrawal, and J. Solà, "Large-scale visual odometry for rough terrain," in *Proceedings of the International Symposium on Robotics Research*, 2007.

[18] I. Parra, M. A. Sotelo, D. F. Llorca, and M. Ocaña, "Robust visual odometry for vehicle localization in urban environments," *Robotica*, vol. 28, no. 3, pp. 441–452, 2010.

[19] A. R. F. Sergio, V. Frémont, and P. Bonnifait, "An experiment of a 3D real-time robust visual odometry for intelligent vehicles," in *Proceedings of the IEEE Conference on Intelligent Transportation Systems (ITSC 09)*, pp. 226–231, 2009.

[20] G. Dubbelman and F. C. A. Groen, "Bias reduction for stereo based motion estimation with applications to large scale visual odometry," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2222–2229, June 2009.

[21] A. I. Comport, E. Malis, and P. Rives, "Real-time quadrifocal visual odometry," *International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 245–266, 2010.

[22] M. Kaess and F. Dellaert, "Visual slam with a multi-camera rig," Tech. Rep. GIT-GVU-06-06, Georgia Institute of Technology, 2006.

[23] Y. S. Chen, L. G. Liou, Y. P. Hung, and C. S. Fuh, "Three-dimensional ego-motion estimation from motion fields observed with multiple cameras," *Pattern Recognition*, vol. 34, no. 8, pp. 1573–1583, 2001.

[24] M. E. El Najjar and P. Bonnifait, "A road-matching method for precise vehicle localization using Belief Theory and Kalman filtering," *Autonomous Robots*, vol. 19, no. 2, pp. 173–191, 2005.

[25] S. Sukkarieh, E. M. Nebot, and H. F. Durrant-Whyte, "A high integrity IMU/GPS navigation loop for autonomous land vehicle applications," *IEEE Transactions on Robotics and Automation*, vol. 15, no. 3, pp. 572–578, 1999.

[26] C. Cappelle, M. E. B. E. Najjar, D. Pomorski, and F. Charpillet, "Intelligent geolocalization in urban areas using global positioning systems, three-dimensional geographic information systems, and vision," *Journal of Intelligent Transportation Systems*, vol. 14, no. 1, pp. 3–12, 2010.

[27] M. Grimes and Y. LeCun, "Efficient off-road localization using visually corrected odometry," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '09)*, pp. 2649–2654, May 2009.

[28] D. A. Forsyth and J. Ponce, *Computer Vision. A Modern Approach*, Prentice Hall, Pearson Education International, 2003.

[29] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of The 4th Alvey Vision Conference*, pp. 147–151, 1988.

[30] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: speeded up robust features," *Proceedings of the ECCV*, vol. 3951, pp. 404–417, 2006.

[31] M. Cazorla, D. Viejo, A. Hernandez, J. Nieto, and E. Nebot, "Large scale egomotion and error analysis with visual features," *Journal of Physical Agents*, vol. 4, no. 1, pp. 19–24, 2010.

[32] F. Solina, "Errors in Stereo due to Quantization," Tech. Rep. MS-CIS-85-34, Department of Computer and Information Science, University of Pennsylvania, 1985.

[33] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pp. 674–679, San Francisco, Calif, USA, 1981.

[34] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Transactions on Pattern Analysis*, vol. 9, no. 5, pp. 698–700, 1987.

[35] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[36] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with apphcatlons to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1981.

[37] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2nd edition, 2004.