An Improved Method of Vehicle Ego-motion Estimation Based on Stereo Vision

Haigen MIN; Zhigang XU; Xiaochi LI; Licheng ZHANG; Xiangmo ZHAO

Department of Traffic Information Engineering & Control, Chang’an University, Xi’an, China, Haigen1990@chd.edu.cn

ABSTRACT: Intelligent vehicle is a senior development form in intelligent transportation system. Acquisition of accurate positioning information with cameras is a core technology in intelligent vehicles. In this paper, we propose a novel method to recover the vehicle trajectory using stereo vision technology. Our method takes advantage of structure and motion-based approaches, so the pose sequence can be computed without prior information on the structure of the relevant scene. In order to improve the efficiency and robustness, a robust feature detection procedure is conducted between image triples. Estimating vehicle ego-motion based on its epipolar geometric constraint makes the presented method not require the time-consuming reconstruction of 3-dimensional scene structure. An improved random sample consensus (RANSAC) algorithm based on geometrical constraints is also employed, which can effectively remove those outliers that are mismatched features or belong to moving objects. For nonlinear problems, an innovative Extended Kalman Filter (EKF) is then adopted to refine the estimated position. Overall, the improvements enable the algorithm to robustly estimate motion in a dynamic environment. Our experiments show that the improved visual odometry (VO) approach performs better than other state-of-the-art positioning methods in terms of computational complexity and accuracy.

Keywords: Visual Odometry; Ego-motion Estimation; Outlier Detection; Extended Kalman Filter

Introduction

The mobile robots should possess ego-motion estimation capabilities in order to roam in their environment. The traditional positioning methods include compass [1], inertial measurement unit [2], wheel odometer [3], GPS [4] and their combinations [5-7], etc. However, there are some disadvantages or limitations of traditional localization methods, e.g., the drift error increases quickly when wheels slip in uneven terrain or other adverse conditions [8]. The degree of GPS accuracy for civilian use may be close to meter while the cost of high precision GPS is too much. What’s more, in GPS-denied environment (e.g., in cities with skyscrapers, forests, tunnels, outer space, etc.), GPS is ineffective because of invalid or absent signals [9]. The integration of those sensors is complicated and extremely costly. The positioning system has rigid requirements as the application environment is becoming complex.

To solve the problems mentioned above, methods based on visual odometry (VO) have received great attentions. Unmanned autonomous vehicle is a hot research topic in the field of intelligent mobile robot. The positioning framework with VO plays a major role in intelligent mobile robot. The research on VO began with Moravec [10], who...
designed a planetary rover equipped with what he termed a slider stereo [11]. On the basis of Moravec’s research, Shafer and Matthies [12-13] employed demonstrated stereo VO’s superiority of accuracy and robustness than monocular VO. Howard [14] implemented a stereo VO with adopted Harris and Fast feature to ensure the real-time performance, and employed feature matching method [15] to find the corresponding feature points. Geiger [16-17] used a simple Sobel template operator to detect feature. They tested the VO algorithm on the KITTI benchmark dataset and obtained the positioning result with high efficiency and accuracy. Jwu-Sheng [18] combined a monocular and an inertial measurement unit (IMU) to form an visual odometer. They employed trifocal tensor geometry information and multi-state constraint Kalman filter in the algorithm architecture to reduce the time consuming and enhance the accuracy of the algorithm. [19] tightly coupled the INS and camera to solve the error accumulation problem of VO, resulting a slow position drift. Besides the fusion applications of visual information and IMU, visual odometry is usually used as supplement to global positioning system (GPS) [20-22] integrated stereo visual-LiDAR odometry and reduced IMU and achieve accuracy at the level of state of art.

This paper proposes an ego-motion estimation method for mobile vehicles, especially ones involving stereo camera rig. The reminder of this manuscript is organized as follows: The system model of our platform, an outline of the overall algorithm, which includes the 2D invariant feature SURF, the outlier removal procedure based on the improved RANSAC algorithm, motion estimation based on the geometric constraint, and the Extended Kalman Filter based refinement, are detailed in algorithm overview. Our experiments and discussion are presented in the experiment. We finish our paper in conclusion.

Algorithm Overview

In this section, we present the overview of our algorithm, mainly including system model, feature detection and matching, outliers’ removal and Kalman Filter procedures. The overall algorithm is depicted in detail with the flow chart (as seen in Fig.1) and following sections.

![Algorithm Overview Diagram](image)

**Figure 1. Sketch of the algorithm architecture.**
It is explained in detail in the flowing sections.
System Model and Parameterization

In general, the stereo camera rig we use in our method can be viewed as a linear pinhole model camera that conforms to the central projection [23-24]. The geometrical relationships of images involve four coordinate systems: the world coordinate system, the camera coordinate system, the physical coordinate system of the image, and the pixel coordinate system. According to the definition of the right-hand spiral rule, the model of the pinhole camera is shown in Fig. 2.

The transformation between point $P$ in the world coordinate system and its projection point $P'$ in the pixel coordinate $(u,v)$ is shown in Eq.(1).

$$
\begin{bmatrix}
 u \\
 v \\
 1
\end{bmatrix} = M_1 M_2 \begin{bmatrix}
 X_w \\
 Y_w \\
 Z_w \\
 1
\end{bmatrix}
$$

(1)

where $M_1$ and $M_2$ are the intrinsic parameter and extrinsic parameter, respectively.

The following Fig.3 shows our experimental platform equipped with a high-resolution stereo camera rig (Basler Ace1600 GigE, image size 800×600 pixels, 20Hz) and differential GPS (NovAtel OEM6TM GNSS). We confirm our claims on data that are captured in a considerably cluttered environment on campus with this driving platform. We use high quality grayscale images for VO we proposed and DGPS data as the ground truth.

Motion parameterization, i.e., determining the spatial position of the camera
coordinate system relative to the world reference system, is expressed by a rotation matrix $[t]$, and a translation vector $R(r)$.

The rotation matrix is defined in Eq.(2) and the rotation angle is parameterized by the Euler angle:

$$R(\theta, \Phi, \Psi) = R_z(\theta) \cdot R_x(\Phi) \cdot R_y(\Psi)$$

(2)

In spatial motion, when the ego-motion vector $(V_x, V_y, V_z, w_x, w_y, w_z)$ of a wheeled car platform and the time difference $\Delta T$ between consecutive frames are known, the values of $t$ and $R(r)$ in each time step can be obtained using (3) and (4). Here, $V_i$ and $w_i$ represent the speed of translation and rotation, respectively.

$$t = (V_x \cdot \Delta T, V_y \cdot \Delta T, V_z \cdot \Delta T)^T$$

(3)

$$R(r) = (w_x \cdot \Delta T, w_y \cdot \Delta T, w_z \cdot \Delta T)$$

(4)

**Feature Detection and Matching**

Feature detection and matching are the critical steps in our algorithm architecture, which restrict the efficiency and precision of the overall algorithm. In this paper, we adopt a robust feature comparing with other 2D invariant features from repeatability, precision-recall and efficiency three aspects, and then we employ circle matching and bucketing concept to refine our matching result.

![Flowchart of the improved SURF algorithm.](image)

We choose a $10\sigma \times 10\sigma$ area and divide it into four subareas to generate a 16-D feature vector rather than the 64-D feature vector described above.

To test the performance of the improved SURF algorithm, we compared it with other feature detection algorithms: specifically SIFT [25], Shi-Tomasi [26], ORB [27] and KAZE [28]. An efficiency function is defined as Eq.(5).
where $N$ is the number of feature points and $T$ is the processing time of the corresponding algorithm. According to Table 1, we conclude that the improved SURF ($P$ achieves the peak) outperforms other feature detection methods.

Table 1. Comparison of five feature detection algorithms.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Improved SURF</th>
<th>SIFT</th>
<th>Shi-Tomasi</th>
<th>ORB</th>
<th>KAZE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection time (T/ms)</td>
<td>12.12</td>
<td>27.32</td>
<td>14.31</td>
<td>10.51</td>
<td>37.46</td>
</tr>
<tr>
<td>Number of feature points (N)</td>
<td>1433</td>
<td>1674</td>
<td>1000</td>
<td>427</td>
<td>626</td>
</tr>
<tr>
<td>Efficiency function ($P$)</td>
<td>59.96</td>
<td>27.17</td>
<td>48.27</td>
<td>57.63</td>
<td>17.19</td>
</tr>
</tbody>
</table>

The improved SURF shows outstanding performance compared to the state-of-the-arts. To demonstrate the excellent performance of this algorithm, we present comparison of experiment results obtained on the evaluation set of [29].

We employ the Euclidean Distance to measure the similarity between SURF descriptors. The K-Nearest-Neighbor (KNN) is adopted to perform the matching procedure between consecutive frames. The ratio test is introduced to remove the mismatches to get a robust feature matching set. A bucketing concept [30] is adopted to choose a subset of the matching feature points. A small number and uniform distribution of feature points reduce the computational complexity of the overall algorithm.

Outliers’ Removal Based on the Improved RANSAC

The result of feature matching in last section contains features of static as well as dynamic objects. In order to maintain the accuracy of subsequent calculations of vehicle position, further elimination of mismatches is needed. The traditional method for the elimination of erroneous matching is the RANSAC algorithm [31]. However, this method requires numerous iterations, which increases computational complexity. Moreover, it often fails to eliminate mismatching. To avoid these shortcomings of those approaches, we propose an improved RANSAC algorithm based on geometric constraints.

There are relationships between two matching points set: (1) the slope of each matching pair is equal or close to each of the others, and (2) the length of each matching pair is equal or close to each of the others. We obtain the set of matching points of the frame $P_1=\{p_{i}(x,y)|p_{i}\in L_1\}$ $(i=0,1,...,n-1)$ and the corresponding set of matching points of the given frame $P_2=\{p_{j}(x,y)|p_{j}\in L_2\}$ $(j=0,1,...,n-1)$, where $n$ is the sum of matching points. A tuple $M(E, C)$ represents a geometric constraints model, i.e., there is a constraint $C=\{c_{j}|j=0,1,...,m\}$ with regard to a finite set of elements $E=\{e_{j}|j=0,1,...,n\}$. To calculate the geometrical relationship between $P_1$ and $P_2$, we assume a point set $P$.

$$P=\{P_1[i],P_2[j]|i=j=0,1,...,n-1\}$$ (6)

Each element of $P$ is a matching point pair. Based on the geometrical constraint model $M$, we search the matching point set $(P_1[i],P_2[j])$ that satisfies the geometric constraints $C$ and reject the matching point pair that does not satisfy it.
Experiments show that our proposed algorithm eliminates mismatching, reduces the number of iterations and improves computational efficiency compared with the traditional RANSAC algorithm. Thus it improves the efficiency of the image matching algorithm to a greater degree, as shown in Table 2.

Table 2. Comparison between conventional RANSAC and the improved RANSAC.

<table>
<thead>
<tr>
<th>Group No.</th>
<th>(N_{\text{ori}})</th>
<th>(P_{\text{correct}}^{\text{ori}}) (%)</th>
<th>(T_{\text{ori}}) (ms)</th>
<th>(N_{\text{our}})</th>
<th>(P_{\text{correct}}^{\text{our}}) (%)</th>
<th>(T_{\text{our}}) (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>245</td>
<td>80.0</td>
<td>29.0</td>
<td>162</td>
<td>100.0</td>
<td>18.3</td>
</tr>
<tr>
<td>2</td>
<td>178</td>
<td>84.6</td>
<td>20.8</td>
<td>112</td>
<td>100.0</td>
<td>12.8</td>
</tr>
<tr>
<td>3</td>
<td>156</td>
<td>84.4</td>
<td>18.9</td>
<td>104</td>
<td>99.9</td>
<td>12.3</td>
</tr>
<tr>
<td>4</td>
<td>99</td>
<td>87.7</td>
<td>11.6</td>
<td>82</td>
<td>100</td>
<td>10.1</td>
</tr>
</tbody>
</table>

In the above table, \(N_{\text{ori}}\) is the number of matching points and \(T_{\text{ori}}\) is the average computation time for the conventional RANSAC. \(N_{\text{our}}\) is the number of matching points and \(T_{\text{our}}\) is the average computation time for the improved RANSAC. The definition of \(P_{\text{correct}}^{\text{ori}}\) and \(P_{\text{correct}}^{\text{our}}\) is as follows: \(N_{\text{correct}}^{\text{ori}}\) is the number of correct matching points for conventional RANSAC, and \(N_{\text{correct}}^{\text{our}}\) is the number of correct matching points for the improved RANSAC.

\[
P_{\text{correct}}^{\text{ori}} = \frac{N_{\text{correct}}^{\text{ori}}}{N_{\text{ori}}} , \quad P_{\text{correct}}^{\text{our}} = \frac{N_{\text{correct}}^{\text{our}}}{N_{\text{our}}} \tag{7}
\]

The improved RANSAC reduces the overall running time, meanwhile enhance the algorithm accuracy.

After the procedure of outlier removal, we get a robust inlier matching set which is the prerequisite of ego-motion estimation. Performing minimization using Eq.(8) obtains the estimation of \([R(r), t]\).

\[
c^2(P_k, P_{k-1}) = \arg\min_{i} \sum_{i=1}^{n_{\text{inl}}} \| P_i - (R(r)P_{k-1} + t) \|^2 \tag{8}
\]

where \(i\) is a feature point. \(k\) is the time instant. \(P_k\) and \(P_{k-1}\) are triangulated 3D points at instants \(k\) and \(k-1\), respectively.

Combining motion parameterization in parameterization part and Eq.(8), we can calculate the initial estimation of 6DOF motion parameters.

**Kalman Filter Based Refinement**

The theory of Kalman Filter[32] is widely applied into dynamic system to estimate the instantaneous state. It provides a solution that may directly reduce the effects of disturbance noises including system and measurement noises. The errors in the parameters can also normally be handled as noise. Using Kalman Filter to estimate the current state of the car, we assumed that measuring value and status are linearly dependent. Kalman filter is a linear filter. Through the state space model of signal and noise, using the previous estimated moment and observed value of the current time to update the current moment state variables and get an estimated value.
State equation: \[ X(k) = AX(k-1) + BU(k) + W(k) \] (9)

Observed equation: \[ Z(k) = HX(k) + V(k) \] (10)

where, \( X(k) \) is the state vector at time \( k \) in the system. \( U(k) \) is the control variate at time \( k \). \( A \) and \( B \) are parameter matrixes of system. \( Z(k) \) represents the measured value at time \( k \) in the system and \( H \) is parameter matrix of the measurement system. \( W(k) \) and \( V(k) \) indicates process noise and observation noise at time \( k \) respectively and \( W(k) \) and \( V(k) \) are the independent zero mean Gaussian white noise sequence. It defines \( Q(K) \) and \( R(K) \) respectively are covariance matrix of \( W(k) \) and \( V(k) \).

\[
E\begin{bmatrix} W(k) \\ V(k) \end{bmatrix} \begin{bmatrix} W'(k) \\ V'(k) \end{bmatrix} = \begin{bmatrix} Q(K) & 0 \\ 0 & R(K) \end{bmatrix}
\] (11)

For pose of every feature point in the previous frame the optimum pose of every feature point in the current frame is calculated. In our system, we set \( R(K) = 1e-2 \) and \( Q(K) = 2e-5 \) to get the optimum result. \( H \) is the unit matrix and system control variate is \( U(k) = 0 \).

**Experiments and Discussion**

To illustrate the advantage of our proposed method, we tested our stereo motion estimation method on the KITTI dataset [17]. The KITTI odometry benchmark consists of 22 stereo sequences and we only randomly took the sequence 07 as comparison data sets.

![Figure 5. Motion estimation on sequence 07 with ground truth, libviso2 stereo VO and our proposed method.](image)
From Fig. 5, we can see that our method is more close to the ground truth than libviso2 stereo method. No matter with regard to rotation error or translation error, our method performs better than libviso2 stereo method presented in Fig. 6.

To further demonstrate that our method is suitable for practical application, we used the vehicle platform presented in system model to capture image data. Our approach was tested with two cameras mounted on top of the vehicle. The distance from the center of the camera lens to the ground was 1.56m. The speed of the vehicle ranged from 0km/h to 40km/h. The sample image frames are shown in Fig. 7. As we can see that some moving targets and GPS-denied environment are captured in the scene.

(a) dynamic targets.  (b) GPS-denied environment.

Figure 7. The sample image frames in our data.

Figure 8. Comparison of the proposed method with other ego-motion estimation methods (DGPS and monocular VO). Left figure shows the trajectory estimated using different methods and right figure shows estimated velocity along x-, y-, and z-axes.
The trajectories recovered using the proposed method and other methods are shown in Fig. 8. The trajectory was approximately 1.3km and lasted 233s. Further, we compared the velocities along the x-, y-, and z-axes, and calculated the Euler angles relative to each axis. We regard the trajectory of DGPS as the ground truth. The proposed method is more close to the ground truth than the stereo method of libviso2. Moreover, our method has less drift than DGPS in a GPS-resistant environment (see Fig. 7(b)), which would ensure a more reliable performance.

Conclusion

In this work, we presented a novel approach for 6DoF ego-motion of stereo VO based on the improved SURF features. A new method for vehicle positioning based on stereo vision is proposed and compared with traditional VO techniques. With our key enhancements to the adopted approach and algorithm, overall the proposed method can generate highly accurate ego-motion estimation results in a manner suited to real-time applications. Extensive research has been carried out in fusion framework. More detailed, we devote to more robust feature detection and matching algorithm, and integrate vision with radar/LiDAR data.

Acknowledgment

This work was supported and made possible by National Natural Science Foundation of China (Grant No. 51278058), Application of Basic Research Project for National Ministry of Transport (Grant No. 2015319812060) and the Fundamental Research Funds for the Central Universities (Grant No. 310824165024).

References


