Vision Navigation Based on the Fusion of Feature-based Methods and Direct Methods

Chuan-qi CHENG*, Xiang-yang HAO, Jian-sheng LI and Xu ZHANG
School of Navigation and Aerospace Engineering, Information Engineering University, China
*Corresponding author

Keywords: Pose estimation, Vision navigation, Feature-based method, Direct method, Lie group, Nonlinear optimization.

Abstract. Vision navigation is an alternative to Global Position System (GPS) in environments where access to GPS is denied. Yet, the classical feature-based method has to extract feature observations from the image as input, which performs poor in environments with little features. To address this problem, a fusion of feature-based method and direct method is designed. The feature-based method is used in feature rich regions, while the direct method is used in regions with little features. Thus we utilize the fusion of two methods to improve the environmental adaptability of vision navigation. To improve the robustness to outliers, the Huber weight function is applied. Then the nonlinear optimization method is utilized to obtain the optimal camera pose. Experimental results demonstrate that the proposed method can meet the needs of real-time autonomous navigation.

Introduction

Vision navigation has become one of the most popular research topics in the field of autonomous navigation. The major reason is its use in robotics and automotive industry, especially to navigate unmanned aerial vehicles (UAVs) in GPS-denied environments. The key technique is estimating the six degrees-of-freedom motion of a camera merely from a stream of images.

In the latest decades, there has been an intense research on vision navigation, which can be divided into two main techniques: feature-based methods [1-5] and direct methods [6-11].

Feature-based methods. The feature-based methods split the overall problem—estimating camera poses from images - into two sequential steps: First, extract a set of feature observations from the images; second, the camera pose is computed as a function of the extracted features only. While the decoupling simplifies the overall problem and can be accurate, robust and reliable in feature-rich environments, it has a limitation: only the information that conforms to the feature type can be used [8], so the large amount of information in images is discarded [12-15]. Thus feature-based methods rely on the set of feature observations and can’t be reliable in feature-less environments.

Direct methods. In contrast to the feature-based method, the direct methods are based on the direct minimization of the photometric pixel values, which don’t extract features and thus can use all the image information [9]. While direct methods outperform feature-based methods in feature-less environments, it was shown that their current accuracy is lower than the feature-based methods [4].

Considering the advantages of both feature-based methods and direct methods, we propose a vision navigation method based on the fusion of the two kinds of methods, which can better adapt all kinds of environments (feature-less or feature-rich) and ensure the accuracy and reliability of vision navigation.

The rest of the paper is organized as follows. Section 2 describes the basic principles of feature-based methods and direct methods. Section 3 presents our vision navigation method in detail. Finally, section 4 shows the experimental results and section 5 presents the conclusions.

Preliminaries

In this chapter we summarize the basic principles of feature-based methods and direct methods.
Feature-based Methods

Feature-based methods are based on the minimization of re-projection error to obtain the optimal camera’s poses [16]. Assume that there are camera represented by their poses \( \{T_i, i = 1, \ldots, m\} \), a set of 3D points in world coordinate frame \( \{y_j, j = 1, \ldots, n\} \) and the image coordinates of \( y_j \) in the camera \( T_i \) is denoted by \( z_{i,j} \).

By using the 3D rigid body transformation \( T_i \), we can map the 3D point \( y_j \) from the world coordinate frame to the camera frame:

\[
x_{i,j} = [I_{3\times3} \ 0_{3\times1}] \cdot T_i \hat{y}_j
\]

where \( x_{i,j} = [x_1, x_2, x_3]^T \) and \( \hat{y}_j = [y_j, 1]^T \) denote the homogeneous coordinate representation.

According to the pinhole camera projection model, we can project \( x_{i,j} \) from the camera frame to the image and obtain the predicted image coordinates \( \hat{z}_{i,j} \):

\[
\hat{z}_{i,j} = \pi(x_{i,j}) = \begin{bmatrix} f_u & 0 & 0 \\ 0 & f_v & 0 \\ x_3 & x_2 & x_1 \end{bmatrix} + \begin{bmatrix} c_u \\ c_v \end{bmatrix}
\]

where \( \pi() \) denotes projection function and \( f_u, f_v, c_u, c_v \) are the camera’s intrinsic parameters.

Thus the re-projection error can be expressed as follows:

\[
e_{i,j}^G = z_{i,j} - \hat{z}_{i,j}
\]

Feature-based methods are based on the minimization of the following cost function to obtain the optimal camera pose:

\[
E = \sum_{j=1}^{n} e_{i,j}^T e_{i,j}^G
\]

Direct Methods

In contrast to feature-based methods, direct methods are based on the direct minimization of the photometric pixel values (photometric error) to obtain camera poses [16]. The photometric error is defined as follows:

\[
e_{i,j}^p(x_j, T_i) = I_i(W(x_j, D_{ref}(x_j), T_i)) - I_{ref}(x_j)
\]

where \( I_i \) denotes the current image, \( I_{ref} \) denotes the reference image, \( T_i \) denotes the pose of current camera, (we assume the reference camera coordinate frame is aligned with the world coordinate frame, so the global pose of current camera is same as the relative pose to the reference camera), \( x_j \) is the image coordinates in the reference image, \( D_{ref}(x_j) \) denotes the inverse depth of \( x_j \), and \( W() \) is the warping function, which is denoted as follows:

\[
W(x_j, D_{ref}(x_j), T_i) = \pi(T_i, \pi^{-1}(x_j, \frac{1}{D_{ref}(x_j)})) = \pi(T_i, y_j) = \pi(x_{i,j})
\]

where \( \pi^{-1}() \) is the inverse of the projection function (formula 2), which is defined as follows:

\[
\pi^{-1}(x, depth_i) = K^{-1} \cdot \begin{bmatrix} x \\ 1 \end{bmatrix} depth_i
\]
where \( K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \), is the calibration matrix of the camera, \( x \) is the 2D image coordinates, and \( depth_i \) is the depth value of \( x \).

The 3D point \( y_j \) in the reference coordinate frame can be computed by follows:

\[
y_j = \pi^{-1}(x_j, \frac{1}{D_{ref}(x_j)}) = K^{-1} \begin{bmatrix} x_j \\ 1 \\ D_{ref}(x_j) \end{bmatrix}
\]

The cost function used in direct methods is:

\[
E = \sum_{j=1}^{n} (e_{i,j}^p)^2
\]

The Fusion of Feature-based Methods and Direct Methods

In this chapter, we present our vision navigation method in detail, including the overview of proposed method, non-linear optimization problem solving, the jacobians in optimization, and some other schemes.

The Overview of Proposed Method

The pipeline of the proposed method is as follows:

Step 1: Initialize pose with motion model. If pose estimation was successful for last frame, we use a constant velocity motion model to predict the camera pose.

Step 2: Feature extraction. We extract FAST corners at different scale levels and compute the orientation and ORB descriptor on the extracted FAST corners [4], to ensure real-time and rotation, scale invariant. If the features are enough, go to step 3, otherwise, go to step 5.

Step 3: Guild search of the 3D points. Based on the initial pose, we can perform a guided search of the 3D points observed in the last frame. If enough correspondences are found, go to step 4, otherwise, go to step 5.

Step 4: 3D-2D pose optimization. Based on the 3D-2D correspondences, we can perform pose optimization via the feature-based method.

Step 5: Pose estimation via direct methods. Based on last frame and the 3D map points, perform the direct method to obtain the camera pose.

Non-linear Optimization Problem Solving

Both feature-based methods and direct method are non-linear optimization problem. The problem is commonly solved via successive linearizations, e.g., the Gauss-Newton (GN) or the Levenberg-Marquardt (LM) methods.

At each iteration, the update in GN method is:

\[
\Delta_k = -(J_k^T J_k)^{-1} J_k^T e_k
\]

The update in GN method is:

\[
\Delta_k = -(J_k^T J_k + \lambda I)^{-1} J_k^T e_k
\]

or

\[
\Delta_k = -(J_k^T J_k + \lambda diag(J_k^T J_k))^{-1} J_k^T e_k
\]

where, \( J = \frac{\partial e}{\partial T} \) is the jacobian matrix, and \( \lambda \) is the damping factor. To ensure robustness and accuracy of vision navigation, we use LM method to solve optimization problem in the paper.
The Jacobians Based on Lie Algebra Pose Representation

The jacobian in minimizing the re-projection residual is as follows.

\[
\mathbf{J}^G_j = \begin{bmatrix}
\frac{f_u}{x_3} & 0 & -\frac{f_u x_1}{x_3^2} \\
0 & \frac{f_v}{x_3} & -\frac{f_v x_2}{x_3^2}
\end{bmatrix}
\begin{bmatrix}
I_{3x3} & -\mathbf{\hat{x}}_{i,j}
\end{bmatrix}
\] (12)

The jacobian of photometric residual in direct methods is expressed as follows:

\[
\mathbf{J}^P_j = \left[ \nabla I_{i,u} \quad \nabla I_{i,v} \right] \begin{bmatrix}
\frac{f_u}{x_3} & 0 & -\frac{f_u x_1}{x_3^2} \\
0 & \frac{f_v}{x_3} & -\frac{f_v x_2}{x_3^2}
\end{bmatrix}
\begin{bmatrix}
I_{3x3} & -\mathbf{\hat{x}}_{i,j}
\end{bmatrix}
\] (13)

The Updating Step and Weight Function

After computing the Jacobian and update at time step k, we can perform updating step as follows:

\[
T_{i,k+1} = \exp(\delta \hat{\mathbf{x}}) T_{i,k}
\] (14)

To ensure the vision navigation method robust to the outliers, a Huber kernel function is applied to weight residuals.

\[
\text{Weight(residual)} = \begin{cases} 
1, & |\text{residual}| < \Delta, \\
\frac{\Delta}{|\text{residual}|}, & \text{otherwise}. 
\end{cases}
\] (15)

where, \(\Delta\) is the predetermined threshold.

Experiments and Analysis

To validate the effectiveness of the proposed method, experiments were performed using the TUM dataset [18]. The TUM dataset consists of 39 sequences that we recorded in two different indoor environments. Each sequence contains the color and depth images captured by Kinect sensor, as well as the ground truth trajectory from the motion capture system. In this chapter we present a comparison of the proposed method and the feature-based method’s performance in vision navigation. All experiments were performed in Ubuntu14.04 system on a Lenovo W520 laptop (i7-2760QM 2.4GHz, RAM 8GB).

To validate the effectiveness of the proposed method, we utilize TUM dataset to perform vision navigation via feature based method and proposed method respectively. Experimental results are shown in Figure 1 and Figure 2.

![Figure 1. Trajectories generated by proposed method and feature based method.](image-url)
As shown in figure 1 and figure 2, in some regions, due to lack of feature information, camera poses obtained via the feature based method have larger errors, while the proposed method are more robust and can generate constant accurate positions of camera, which validates the effectiveness of proposed method.

Conclusions
This paper presents a vision navigation method based on the fusion of feature based method and direct method to improve the robustness and accuracy of vision navigation. Experiments on TUM dataset validate the effectiveness of proposed method and the main conclusions are as follows:

1) Compared with the feature based method, the proposed method in this paper is more adaptable to the environment and is more suitable for autonomous navigation of unmanned systems.

2) Vision navigation is essentially one way of dead reckoning and it is inevitable that the error will accumulate. To guarantee the effectiveness of long-term autonomous navigation, we will study the fusion of multi-sensor further.

Reference


