IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Study on the method of SLAM initialization for monocular vision

Study on the method of SLAM initialization for monocular vision

Guanghui Xu[1], Qingsong Zhang[1*], Na Li[1]

[1] College of Communication Engineering, Army University of Science and Technology, Nanjing, Jiangsu, 210007, China

[*]Corresponding author's e-mail:853868426@qq.com

Abstract. Visual SLAM is considered as a key technology to realize the autonomous positioning and navigation of mobile robots, and also a hot research technology in the fields of unmanned driving, augmented reality and smart home. Among them, monocular camera based visual SLAM is one of the hotspots in the field of visual SLAM. In today's mainstream visual SLAM, PTAM needs to manually select two images to complete the initial keyframe trajectory and map point estimation process, which limits the practical application of the system and reduces the success rate of initialization. In addition, ORB_SLAM system adopts the statistical model selection method to realize the automatic initialization process. This paper first introduces the definition of the homography matrix and the fundamental matrix and the implementation algorithm, and then elaborates on the use of a certain scoring strategy to select the model method for initialization.

1. Introduction

Simultaneous Localization and Mapping (SLAM) is based on the Simultaneous Localization and Mapping of the information of the unknown environment perceived by the sensor, the motion trajectory of the sensor in the unknown environment is estimated, and the structural map of the environment is constructed simultaneously [1]. The goal of monocular vision SLAM map initialization is to construct the initial 3d map points and the initial key frame pose. Since the relative depth information cannot be obtained from a single frame image, it is necessary to select two or more frames of images from the image sequence, estimate the camera posture and reconstruct the initial 3d map points [2].

There are two common ways to initialize a map. The first method is based on the assumption that there is a plane object in the space, and two images at different positions are selected to estimate the pose by calculating the homography matrix [3]. The second method is to calculate the basic matrix based on the feature point matching between two frames and further estimate the pose [4]. This method requires the existence of non-coplanar feature points. Both methods have their own limited scenarios, and Mur-Artal proposed a model selection method based on statistics. This method gives priority to the second method and expects to automatically select the first method in the case of scene degradation [5]. If the selected two frames do not meet the requirements, the two frames will be abandoned and re-initialized, which is also the initialization method to be studied in this paper.

The flow chart of the model selection automatic initialization method based on statistics is shown in figure 1. The first step of maping initialization is to calculate the homography matrix H and the fundamental matrix F. First, the homography matrix is calculated from two pre-processed images, which can be generally realized by using the normalized direct linear transformation algorithm. In addition, the normalized eight-point algorithm can be used to calculate the fundamental matrix. The re-projection error of each point is calculated by the homography matrix and the fundamental matrix, and the corresponding chi-square distribution is compared. The total score of the interior points corresponding to the two matrices is calculated respectively, and then the fundamental matrix or the homography matrix or neither is selected according to a certain selection model. Since each calculation of single or fundamental matrix only uses 8 point pairs at most, there will be great uncertainty if it is only calculated once. Therefore, the RANSAC method is introduced to calculate multiple times and further eliminate the outer point.

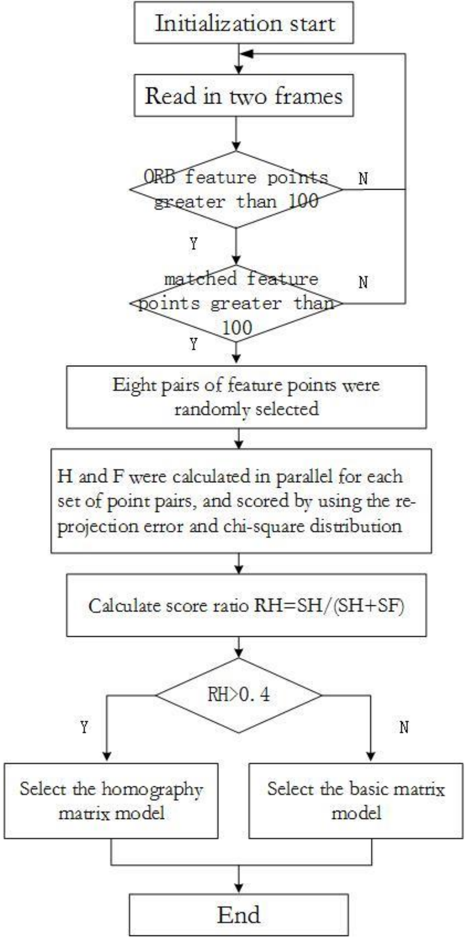
Initialization start
Read in two frames
ORB feature points greater than 100
N
Y
matched feature points greater than 100
N
Y
Eight pairs of feature points were randomly selected
H and F were calculated in parallel for each set of point pairs, and scored by using the re-projection error and chi-square distribution
Calculate score ratio RH=SH/(SH+SF)
RH>0. 4
Y
N
Select the homography matrix model
Select the basic matrix model
End

Fig.1 Map initialization method flow chart

2. Homography matrix model

2.1. Homography matrix

Homography matrix describes the correspondence between two images of three dimensional points on the same plane in space. What needs to be emphasized here is the same plane, as shown in figure 2. The homography matrix can be applied to image correction, image registration, Angle conversion and the calculation of camera motion (rotation and translation) of two images.

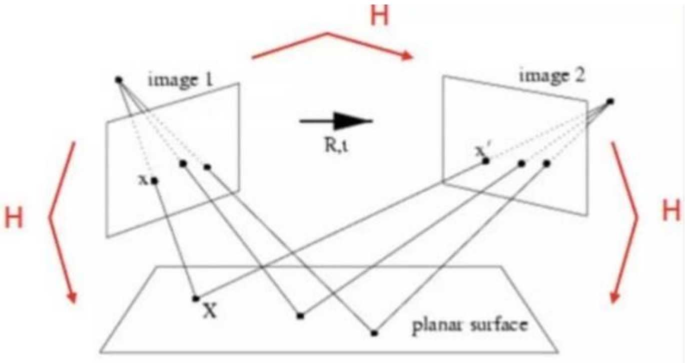

Fig.2 The same plane from different angles

By relying on the basic principle of camera imaging, we can obtain the transformation from the world coordinate system to the camera coordinate system, and the following formula is obtained:

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \frac{1}{Z}\begin{pmatrix} f_1 & 0 & c_x \\ 0 & f_2 & c_y \\ 0 & 0 & 1 \end{pmatrix}\begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix}\begin{pmatrix} x_b \\ y_b \\ z_b \\ 1 \end{pmatrix} = \frac{1}{Z}KT\begin{pmatrix} x_b \\ y_b \\ z_b \\ 1 \end{pmatrix} \tag{1}$$

Where, the point P_X (u,v,1) is the pixel coordinate in the image coordinate system, and the point P_y (x_b , y_b , z_b , 1) is the common coordinate in the world coordinate system. And, f_1 , f_2 is the focal length in the x and y directions, which is generally the same , c_x , c_y is the position of the optical center, which is generally half the length and width, and they are both called the internal parameter K, which is the internal parameter matrix. R is the rotation matrix and t is the translation vector, and they are written together as a matrix in the form T, which is the external parameter matrix, representing the transformation from the world coordinate system to the camera coordinate system.

At this time, we simplify equation (1) and use P to represent the middle part, and get:

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = p\begin{pmatrix} x_b \\ y_b \\ z_b \\ 1 \end{pmatrix} \tag{2}$$

At this point, the matrix P can be regarded as a 4×4 matrix. If the spatial points are in the same plane, we can consider z_b as 0, so the P matrix becomes a 3×3 matrix. For two different cameras, the pixel coordinates and spatial point coordinates can be written as follows:

$$\begin{pmatrix} u_1 \\ v_1 \\ 1 \end{pmatrix} = p_1\begin{pmatrix} x_b \\ y_b \\ 1 \end{pmatrix} \tag{3}$$

$$\begin{pmatrix} u_2 \\ v_2 \\ 1 \end{pmatrix} = p_2\begin{pmatrix} x_b \\ y_b \\ 1 \end{pmatrix} \tag{4}$$

So let's combine these two to get this:

$$\begin{pmatrix} u_2 \\ v_2 \\ 1 \end{pmatrix} = p_2 p_1^{-1} \begin{pmatrix} u_1 \\ v_1 \\ 1 \end{pmatrix} = H \begin{pmatrix} u_1 \\ v_1 \\ 1 \end{pmatrix} \quad （5）$$

H is the homography matrix. On both sides of the H matrix are matching point pairs corresponding to the two images. In other words, the homography matrix H maps the points of the same plane in the three-dimensional space to the imaging image coordinates of the two cameras.

2.2. Calculation of homography matrix

If the adjacent two images are induced by a certain plane in the space, then the corresponding computing homography matrix can be obtained through the 2D of the two images. X_i is the point set on the first image and the point set on the corresponding second image is X_i'. Then the homography matrix $HX_i = X_i'$ is a 3×3 matrix H. This is an equation involving homogeneous vectors, they have the same direction but different scaling factors, and this equation can be expressed in terms of $HX_i \times X_i' = 0$. In this form we can derive a simple linear solution to H.

Assuming $X_i' =（x_i', y_i', z_i'）$, so that can be given explicitly:

$$HX_i \times X_i' = \begin{bmatrix} y_i' h^{3T} X_i - w_i' h^{2T} X_i \\ w_i' h^{1T} X_i - x_i' h^{3T} X_i \\ x_i' h^{2T} X_i - y_i' h^{1T} X_i \end{bmatrix} = 0 \quad （6）$$

Further, formula (6) can be simplified as follows:

$$\begin{bmatrix} 0^T & -w_i' X_i^T & y_i' X_i^T \\ w_i' X_i^T & 0^T & -x_i' X_i^T \\ -y_i' X_i^T & x_i' X_i^T & 0^T \end{bmatrix} \begin{bmatrix} h^1 \\ h^2 \\ h^3 \end{bmatrix} = 0 \quad （7）$$

These expressions have the following form $B_i h = 0$, where B_i is a 3×9 matrix, h is a 9-dimensional vector composed of the elements of the matrix H, and the formula is as follows:

$$h = \begin{bmatrix} h^1 \\ h^2 \\ h^3 \end{bmatrix} \; H = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix} \quad （8）$$

From the above derivation, the homography matrix can be obtained directly from the normalized direct linear transformation algorithm. The flow chart of the algorithm is shown in figure 3.

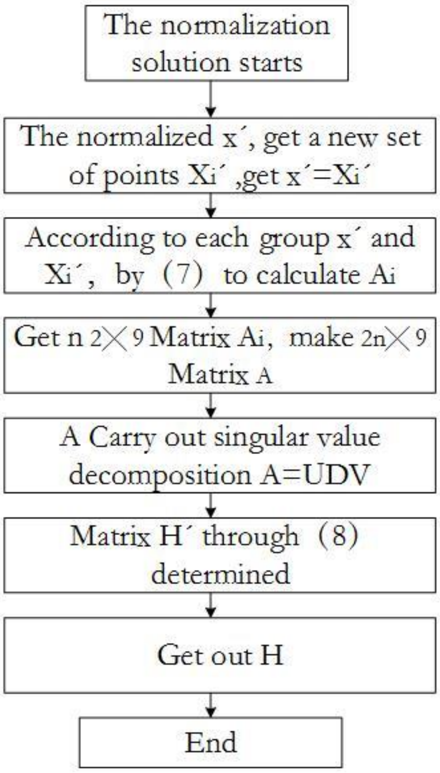
The normalization solution starts
The normalized x´, get a new set of points Xi´ ,get x´=Xi´
According to each group x´ and Xi´, by（7）to calculate Ai
Get n 2╳9 Matrix Ai, make 2n╳9 Matrix A
A Carry out singular value decomposition A=UDV
Matrix H´ through（8）determined
Get out H
End

Fig.3 Algorithm flow chart for solving homography matrix

3. Fundamental Matrix model

3.1. Fundamental Matrix

The correspondence between the image points and polar lines described by polar geometry can be represented by the Fundamental Matrix. In other words, the Fundamental Matrix is an algebraic representation of polar geometry.

It is assumed that the two camera matrices are Q, Q', and denotes that the image plane of the two cameras is U, U', then the parametric equation of the inverse projection line of $\forall \mathrm{m} \in U$ is:

$$Y(x) = Q^* m + xC, s \in (-\infty, +\infty) \qquad (9)$$

Where, Q^* is the generalized inverse of Q, C is the optical center of one of the cameras. Thus, the expression of the fundamental matrix between two cameras can be derived:

$$F = (Q^* C) Q' Q^* m \qquad (10)$$

3.2. Calculation of fundamental matrix

The fundamental matrix is defined by the following equation:

$$x'Fx = 0 \qquad (11)$$

Where x' and x is the Any pair of points in two images. At this time, as long as enough matching point pairs are selected, the basic matrix F can be calculated. If n groups of points are matched, the linear equations are as follows:

$$Af = \begin{pmatrix} x_1 x_1' & x_1' y_1 & x_1' & y_1 x_1' & y_1' y_1 & y_1' & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n x_n' & x_n y_n' & x_n' & y_n' x_n & y_n' y_n & y_n' & x_n & y_n & 1 \end{pmatrix} f = 0 \qquad (12)$$

From the above derivation, the normalized 8-point algorithm can be obtained, as shown in figure 4.

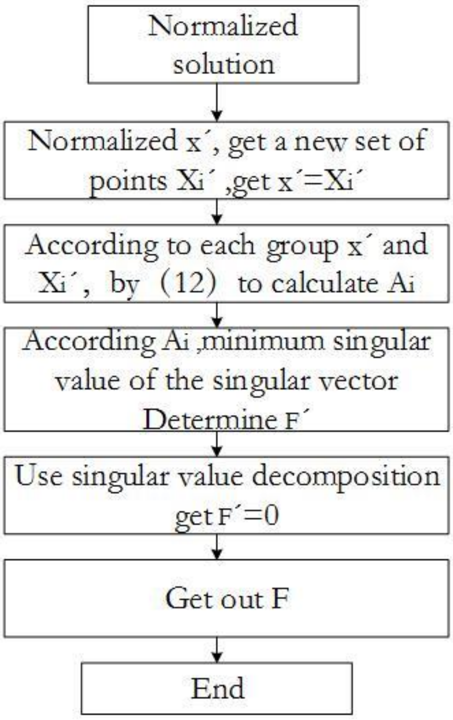
Normalized solution
Normalized x´, get a new set of points Xi´ ,get x´=Xi´
According to each group x´ and Xi´, by （12） to calculate Ai
According Ai ,minimum singular value of the singular vector Determine F´
Use singular value decomposition get F´=0
Get out F
End

Fig.4 Algorithm flow chart for solving the fundamental matrix

4. Choose the model according to the score

4.1. RANSAC algorithm detailed solution

Since each calculation of single or fundamental matrix only uses 8 point pairs at most, there will be great uncertainty if it is only calculated once. Therefore, the RANSAC method is introduced to calculate multiple times and further eliminate the outer point.

The input of RANSAC algorithm is a set of observation data (often containing large noise or invalid points), a parameterized model used to interpret the observation data and some trusted parameters. RANSAC achieves this goal by repeatedly selecting a random subset of the data. The selected subset is assumed to be a local point and verified by the following methods:

1. A model is adapted to the assumed local point, that is, all unknown parameters can be calculated from the assumed local point.

2. Test all other data with the model obtained in 1. If a point applies to the estimated model, it is considered as a local point.

3. If enough points are classified as assumed local points, the estimated model is reasonable enough.

4. Then, the model is reestimated with all the assumed local points (for example, using the least square method), because it has only been estimated with the initial assumed local points.

Finally, the model is evaluated by estimating the error rate of local point and model.

The above process is repeated a fixed number of times, each resulting model is either discarded because there are too few local points or chosen because it is better than the existing model.

4.2. Model selection

In order to obtain more accurate data, the method of multiple RANSAC iterations is usually used to calculate the homography matrix H and the fundamental matrix F, and score the results this time. Keep the H and F with the highest score, and choose F or H according to certain rules to restore the initial relative posture or neither. In each iteration, calculated for each model M(M is homography matrix model H or fundamental matrix model F). The calculation formula is as follows:

$$S_m = \sum_i (p_m(d_{cr}(x_c^{\ i}, x_r^{\ i}, M)) + p_m(d_{rc}(x_r^{\ i}, x_c^{\ i}, M)))$$
$$p_m = \begin{cases} Q - d^2, d^2 < T_m \\ 0, d^2 \geq T_m \end{cases} \tag{13}$$

Where, x_c x_r are the matching point pairs in two frames of images respectively, and $d_{cr}{}^2$, $d_{rc}{}^2$ is the re-projection error from one frame to another. T_m is the re-projection threshold of the outer point when the standard deviation of the measurement error is 1 pixel and 95% is the chi-square test of the inner point. When m is H, T_H =5.99; When M is F, T_F =3.84. Q is defined as T_H , so that the scores of the two models are relative to the same size of the inner point area, which makes this method more fair. When neither model has enough interior points, the current initialization is abandoned and the initialization is restarted.

The scoring rules above are based on the assumption that the probability distribution of the distance between the interior point and the model is the chi-square distribution. The chi-square distribution shows the following relation:

$$\begin{cases} \text{Interior point}, d^2 < \mathrm{t}^2 \\ \textit{exterior point}, d^2 \geq \mathrm{t}^2 \end{cases} \quad \text{and} \quad \mathrm{t}^2 = F_m{}^{-1}(a)\varepsilon^2 \tag{14}$$

For the basic matrix model, the value of t^2 is $3.84\varepsilon^2$; When the single response model, the value of t^2 is $5.99\varepsilon^2$.

If the scene perceived by monocular is flat or approximately flat, it should be explained by 2D monocular. In this case, the fundamental matrix can also be calculated, but the constraint of the opposite pole cannot be well satisfied. At this time, any attempt to restore the motion from the fundamental matrix will produce wrong results, and the homography matrix should be selected to properly initialize from the plane, or the initialization should be abandoned when low parallax is detected. On the other hand, when the scene is non-planar and has sufficient parallax, then the fundamental matrix should be used to explain. In this case, when some point pairs are on the same plane, the homography matrix can also be calculated, but in this case, the homography matrix to restore the motion is not the best choice, but the fundamental matrix should be selected. Here is a robust test formula based on the score:

$$R_H = \frac{S_H}{S_H + S_F} \tag{15}$$

When R_H >0.40, it indicates that the scene belongs to plane, approximate plane or low parallax, and the homography matrix is selected. Otherwise select the fundamental matrix.

After that, through the sports recovery Structure (Structure from Motion, SfM) test, the characteristics of the input image feature matching, homographic matrix or basic matrix is used to estimate the camera pose and map point location, and then using a known as the bundle set adjustment (bundle adjustment, BA) of the nonlinear optimization method to minimize the heavy projection error, thereby restoring the scenario [6-7].

5. Conclusion

This paper mainly studies the initialization method of monocular vision SLAM system. In the current mainstream mode, the basic matrix is usually preferred to estimate the camera pose and map point position, and the homography matrix is automatically used to estimate the camera pose and map point position when the scene degenerates into a plane. Firstly, the definition and algorithm of homography matrix and fundamental matrix are introduced, and then some scoring strategies are used to select the initial model, so as to realize the automatic initialization of monocular vision SLAM system. This method enhances the robustness and usability of the system, making SLAM system more suitable for practical application.

References

[1] Leonard, J., Durrant-Whyte, H. (1991)Mobile Robot Localization by Tracking Geometric Beacons .J .IEEE Transactions on Robotics and Automation., 7:376-382.
[2] Davide, S., Friedrich, F. (2011)Visual Odometry, Part I: The First 30 Years and Fundamentals .J. IEEE Robotics&Automation Magazine., 2011:80-92.
[3] Faugeras O.D., Lustnian, F. (1988)Motion and structure from motion in a piecewise planar environment.J. International Journal of Pattern Recognition and Artificial Intelligence., 2: 485-508.
[4] Hartley, R., Zisserman, A. (2003) Multiple view geometry in computer vision. Cambridge university press, America.
[5] Mur-Artal, R., Montiel, M., Tardos, J.D. (2015)ORB_SLAM: A versatile and accurate monocular SLAM system. J. IEEE Transactions on Robotics., 31: 1147-1163.
[6] Klein, G., Murray, D. (2007)Parallel tracking and mapping for small AR workspaces CJ//Mixed and Augmented Reality. In: 6th IEEE and ACM International Symposium. America. pp. 225-234.
[7] Strasdat, H., Davison, A,J. (2012) Visual SLAM: why filter? .J. Image and Vision., 30: 65-77.