Sensors and Materials, Vol. 28, No. 6 (2016) 667–674
MYU Tokyo

S & M 1223

Determination of Map Scale and Initialization of Landmarks for Aerial Robot Monocular Visual Localization and Mapping

Yin-Tien Wang,[*] Chung-Hsun Sun, and Ting-Wei Chen

Department of Mechanical and Electro-Mechanical Engineering, Tamkang University,
New Taipei City 251, Taiwan

(Received March 3, 2015; accepted February 29, 2016)

Keywords: visual simultaneous localization and mapping (vSLAM), image depth parameterization, data association, map management.

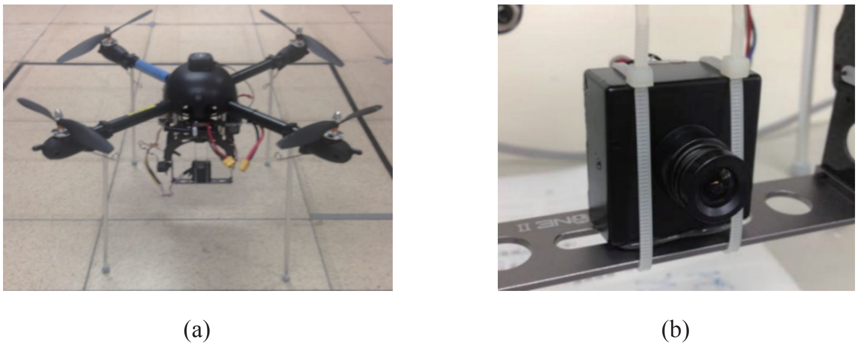
This study investigates the issues of visual-sensor-assisted aerial robot navigation. The major objectives are to give an aerial robot the capabilities of localization and mapping in global positioning system (GPS) denied environments. When an aerial robot navigates in a GPS-denied environment, the visual sensor could provide measurements for estimation of the robot's state and environmental mapping. Considering the carrying capacity of an aerial robot, a single camera is used in this study and the image is transmitted to a PC-based controller for image processing using a radio frequency module. An extended Kalman filter is used as the state estimator to recursively predict and update the states of the aerial robot and the environmental landmarks. The contributions of this study are twofold: First, an ultrasonic sensor is used to provide one-dimensional distance measurements and solve the map scale determination problem of monocular vision. Second, the image depth is represented using the inverse depth parameterization method and initialization of the image features is achieved by a non-delayed procedure. The software program of the robot navigation system was developed in a PC-based controller. The navigation system integrates the sensor inputs, image processing, and state estimation. The resultant system was used to perform the tasks of simultaneous localization and mapping for aerial robots.

1. Introduction

When an aerial robot is navigating in an unknown environment, it relies on sensors to recognize the outside world and then estimate the state of the robot itself to achieve the task of autonomous navigation. Commonly used sensors include the laser range finder (LRF), global positioning system (GPS), and vision sensor. A LRF can offer data from high-precision measurements, but it is too expensive to be extensively used. A GPS signal is free to access, but it is not available for robot navigation in indoor environments. A vision sensor has a reasonable cost and is generally used as a robot's sensing device, especially in a GPS-denied environment. Considering the carrying capacity of an aerial robot, a single camera is used in this study, as shown in Fig. 1. The monocular vision sensor captures only two-dimensional images and lacks depth information on environmental objects. Without depth information, the spatial coordinates of a new landmark cannot be determined. In addition, the map scale of the environment cannot be initially estimated.

[*]Corresponding author: e-mail: ytwang@mail.tku.edu.tw

ISSN 0914-4935 © MYU K.K.



(a) (b)

Fig. 1. (Color online) (a) Quadrotor aerial robot and (b) monocular vision sensor.

Many researchers have developed landmark initialization procedures either in a time-delayed method or an un-delayed method for monocular vision.(1,2) The un-delayed method was utilized in this research. When an image feature is selected, the spatial coordinates of the image feature are calculated by employing the method of inverse depth parameterization.(2) However, the problem of determining the map scale remains unsolved. In this study, an ultrasonic sensing system was developed to provide one-dimensional distance measurement and solve the map scale determination problem of monocular vision.

The features of images detected by the vision sensor can be used to represent the landmarks in an environment and build an environmental map for aerial robot navigation. A detection method based on the scale-invariant feature was developed by Lindeberg.(3) An image feature is selected by examining the determinant of the Hessian matrix based on the non-maximum suppression rule. Scale-invariant features have the advantages of high stability and repeatability; however, they have the disadvantage of extensive computation. Concerning the issue of computational speed, Bay *et al.* replaced the Gaussian second-order derivative with the box filter and calculated the approximation of the determinant of the Hessian matrix using the integral image method.(4) This method, called speeded-up robust features (SURFs), significantly reduces calculation times. In this study, the SURF algorithm was employed to detect the features from monocular RGB images and to represent the landmarks in the environmental map. Meanwhile, an extended Kalman filter (EKF) was used to recursively predict and estimate the robot's state as well as the environmental landmarks.(5)

The contributions of this paper are the novel procedures used to solve the problem of determining the map scale as well as initializing new landmarks for monocular vision in robot navigation. In this study, we also extended the usability of local invariant feature detectors in the tasks of simultaneous localization and mapping (SLAM) by utilizing SURFs' robust representation of visual landmarks. Data association and map management for SURF-based mapping were also developed to improve the robustness of SLAM systems.

The paper is organized as follows: Section 2 presents the structure of the aerial robot SLAM. Monocular visual measurement and the method of SURF-based mapping are described in §§ 3 and 4, respectively. Section 5 presents the experimental applications and results. Finally, concluding remarks are given in the last section.

2. Aerial Robot SLAM

When an aerial robot performs SLAM tasks, the states of the robot and landmarks in the environment are estimated on the basis of measurements. In this study, a monocular vision system was used as the only measuring device in the state estimation algorithm. The monocular camera was carried by the aerial robot and was treated as a free-moving system with unknown inputs.(1) System states were estimated using an EKF estimator to solve the target tracking problem.(1,6) The state sequence of a system at time step k can be expressed as

$$\boldsymbol{x}_k = f(\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1}, w_{k-1}), \tag{1}$$

where $\boldsymbol{x}_k$ is the state vector, $\boldsymbol{u}_k$ is the input, and w_k is the process noise. When performing SLAM tasks using a vision sensor, the state vector contains the states of the robot and landmarks,

$$\boldsymbol{x} = [\boldsymbol{x}_{\mathrm{C}}^{\mathrm{T}}, \boldsymbol{M}^{\mathrm{T}}]^{\mathrm{T}} = [\boldsymbol{x}_{\mathrm{C}}^{\mathrm{T}}, \boldsymbol{m}_1^{\mathrm{T}}, \boldsymbol{m}_2^{\mathrm{T}}, ..., \boldsymbol{m}_j^{\mathrm{T}}]^{\mathrm{T}}, \tag{2}$$

where $\boldsymbol{x}_{\mathrm{C}} = [\boldsymbol{r}^{\mathrm{T}}, \boldsymbol{\phi}^{\mathrm{T}}, \boldsymbol{v}^{\mathrm{T}}, \boldsymbol{\omega}^{\mathrm{T}}]^{\mathrm{T}}$ denotes the robot's position and velocity in the world frame, and $\boldsymbol{m}_j$ represents the jth landmark in the environment map $\boldsymbol{M}$. The objective of the robot's SLAM tasks was to estimate the state $\boldsymbol{x}_k$ of the target recursively according to the measurement $\boldsymbol{z}_k$ at k,

$$\boldsymbol{z}_k = g(\boldsymbol{x}_k, v_k), \tag{3}$$

where v_k is the measurement noise. Since the sensor frame was set at the center of the camera, the coordinates of ith observed image feature in the world frame (Fig. 2) were

$$\boldsymbol{m}_i = \boldsymbol{r} + \boldsymbol{h}_i^W = \boldsymbol{r} + \boldsymbol{R}\boldsymbol{h}_i^C, \tag{4}$$

where $\boldsymbol{r}$ is the position vector of the sensor frame, $\boldsymbol{R}$ is the rotational matrix(7) from the world frame to the sensor frame, and $\boldsymbol{h}_i^W$ and $\boldsymbol{h}_i^C$ are the ray vectors of ith image feature in the world and sensor frames, respectively. Because of the lack of one-dimensional range information in monocular vision, how to initialize the image features as new landmarks appear becomes an important topic. In this study, a visual landmark initialization procedure based on the inverse depth parameterization(2) was developed and is described in the following section.

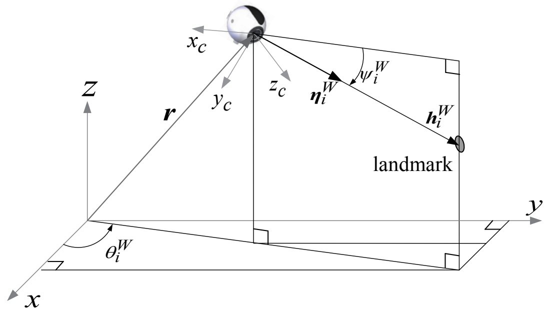


Fig. 2. Coordinates of monocular vision sensor system.

3. Monocular Visual Measurement

3.1 Inverse depth parameterization

For the initialization of new landmarks in the monocular vision system, the un-delayed method was used in this research. When an image feature was selected, the spatial coordinates of the image feature were calculated by employing the method of inverse depth parameterization.(2) Assume that there are n image features with position vectors, $\boldsymbol{m}_i$, $i = 1, \dots, n$, which are described by the 6-dimentional (6D) state vectors

$$\hat{\boldsymbol{m}}_i = [\hat{r}^W_{ix}\ \hat{r}^W_{iy}\ \hat{r}^W_{iz}\ \hat{\theta}^W_i\ \hat{\psi}^W_i\ \hat{\rho}_i]^{\mathrm{T}}, \tag{5}$$

$\hat{\boldsymbol{r}}^W = [\hat{r}^W_{ix}\ \hat{r}^W_{iy}\ \hat{r}^W_{iz}]^{\mathrm{T}}$ indicates the estimated state of the camera when the feature was observed, as shown in Fig. 2; $\hat{\rho}_i$ is the estimated image depth of the feature; and $\hat{\theta}^W_i$ and $\hat{\psi}^W_i$ are the longitude and latitude angles of the spherical coordinate system which is located at the camera center. To compute the longitude and latitude angles, a normalized vector $\boldsymbol{\eta}^W_i$ in the direction of the ray vector was constructed using the perspective project method:

$$\boldsymbol{\eta}^W_i = \boldsymbol{R}(\hat{\phi}^W)\left[\frac{I_{ix} - u_0}{f_u}\ \ \frac{I_{iy} - v_0}{f_v}\ \ 1\right]^{\mathrm{T}}. \tag{6}$$

Focal lengths f_u and f_v denote the distance from the camera center to the image plane in the u- and v-axes, respectively; (u_0, v_0) is the offset pixel vector of the image plane; and (I_{ix}, I_{iy}) are the image coordinates of the feature. Therefore, from Fig. 2, the longitude and latitude angles of the spherical coordinate system can be obtained as

$$\hat{\theta}^W_i = \tan^{-1}\left(\frac{\eta^W_{iy}}{\eta^W_{ix}}\right), \tag{7}$$

$$\hat{\psi}^W_i = \tan^{-1}\frac{\eta^W_{iz}}{\sqrt{(\eta^W_{ix})^2 + (\eta^W_{iy})^2}}. \tag{8}$$

When image features were selected to be new landmarks, the inverse depth parameterization vector in Eq. (5) was assigned to be new augmented states in the EKF-based SLAM. However, the inverse depth coordinates are 6D state vectors and computationally costly. A switching criterion was established in reference 2 based on a linearity index. If the linearity index is satisfied, then the 3D state vector in Eq. (4) is modified to replace the 6D state vector as:

$$\hat{\boldsymbol{m}}_i = \begin{bmatrix} \hat{r}^W_{ix} \\ \hat{r}^W_{iy} \\ \hat{r}^W_{iz} \end{bmatrix} + \frac{1}{\hat{\rho}_i}\hat{\boldsymbol{\eta}}(\hat{\theta}^W_i, \hat{\psi}^W_i), \tag{9}$$

where $\hat{\boldsymbol{\eta}}(\hat{\theta}^W_i, \hat{\psi}^W_i)$ is the unit ray vector computed from the estimated states.

3.2 Determination of map scale

To determine the map scale in a monocular SLAM problem, we developed a one-dimensional distance detector based on ultrasound technology. The distance detector consists of an ultrasound sensor chip (HC-SR04), a radio frequency transmitter (3Dr Telemetry), and a microchip (Arduino Nano). The circuit board of the detector is shown in Fig. 3. When the aerial robot is taking off, the ultrasound sensor is designed to measure the distance from the ground. The SLAM task begins to work if the height of the quadrotor is 1.5 m above the ground. At the beginning of the SLAM task, some SURF features obtained from the first image were chosen as the map landmarks and their states were initialized according to Eq. (4). In the equation, the depth information of these SURF features is obtained from the ultrasound sensor. From these initial SURF features, the map scale was also calculated. After the map scale was obtained, the ultrasound sensor was turned off and the newly added landmarks were initialized using the 6D state vector in Eq. (5).

4. SURF-Based Mapping

Robot visual mapping requires a robust method of representing visual landmarks detected in an image. In this study, we used the SURF method to detect and represent visual landmarks for robot mapping during SLAM tasks. The SURF method developed by Bay *et al*. uses a box filter instead of a difference of Gaussians to approximate the determinant of the Hessian matrix.(4) The box filter was further combined with the integral image method to reduce the image processing time.(8) After the features were detected from the image, the description vector was computed to represent feature characteristics. A high-dimensional description vector was used to describe the uniqueness of the feature. For matching high-dimensional description vectors, the most popular method is the nearest-neighbor search method.(9) The criterion for matching two image features is usually to determine the smallest Euclidian distance between their descriptors.

To implement the navigational tasks, the monocular vision was integrated with the free-moving motion model, the measurement model, and the SURF detection algorithm to form a SLAM system. Once the images were captured by the camera, image features were detected using the

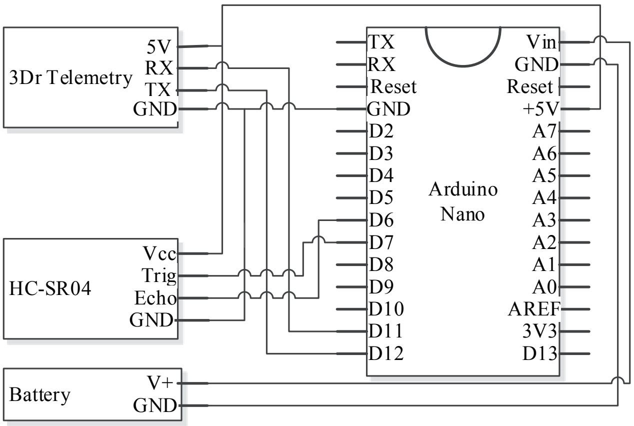


Fig. 3. Circuit diagram of the ultrasound distance detector.

SURF method. The system performed data association of the map landmarks with the image features using the proposed matching criterion. A map management system was also designed to coordinate the newly added features and the “bad” features in the system. New features were chosen as landmarks and added to the map when the robot explored an unknown environment. The state variables of all new landmarks were augmented in the state vector in Eq. (1). However, features that were not continuously detected during the task were considered “bad” features and were deleted from the state vector.

5. Experimental Results

The SLAM experiment was implemented on a quadrotor aerial robot to validate the proposed algorithms. In this experiment, the monocular vision was carried by the quadrotor to follow a forward-backward trajectory of 4 m length, as shown in Fig. 4. The camera lens always faced downward as the aerial robot flew along the forward-backward trajectory 2.5 times. The computer used in the experiments was an ASUS X550V notebook with Intel Core i5-3230 CPU at 2.6 GHz and 8 Gb RAM. The SLAM system started when the quadrotor’s height was 1.5 m as measured by the ultrasound sensor. Fourteen SURF features from the first image were chosen as map landmarks and their state vector was initialized according to Eq. (4) in which the image depth information was obtained from the ultrasound sensor. After that, new landmarks were added to the map constantly and their state vectors were initialized using Eq. (5). As the monocular vision followed the forward-backward trajectory, the SLAM system concurrently built the environment map and estimated the robot’s pose. Figure 5 shows the 3000th image frame obtained in the experiments; the captured RGB image is shown in the left panel. Thirteen landmarks were detected at this frame and the map size was increased to 170. The top-view (*xy*-plane) and side-view (*yz*-plane) plots of the environmental map are shown in the middle and right panels, respectively. The map size and sampling frequency versus the image frame is plotted in Fig. 6. The map size was increased to be 140 landmarks at the end of the first forward-backward trip, to be about 160 landmarks at the end of the second trip, and to be 176 landmarks at the end of the experiment. The average sampling frequency was about 15 Hz and the lowest frequency was about 7 Hz. Figure 7 shows the deviations of the robot’s pose estimation in the xyz-axes. The figure shows that, during the SLAM task, the average pose deviation was less than 10 cm, and the highest peak was about 20 cm.

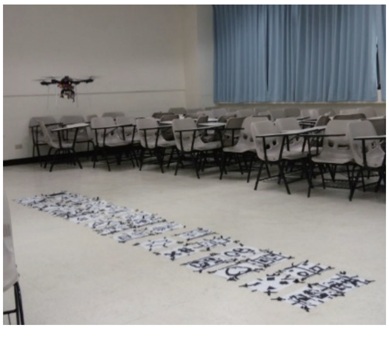


Fig. 4. (Color online) Trajectory of quadrotor SLAM task.

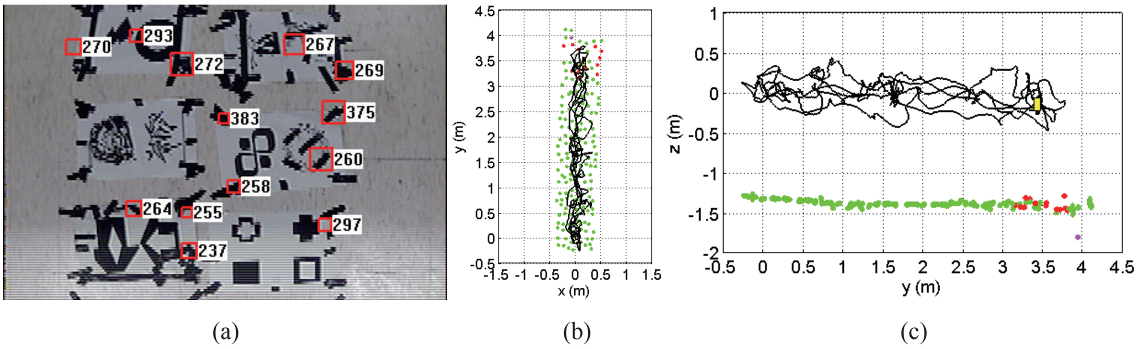


Fig. 5. (Color online) (a) Image, (b) top-view map (*xy*-plane), and (c) side-view map (*yz*-plane) of 3000th frame.

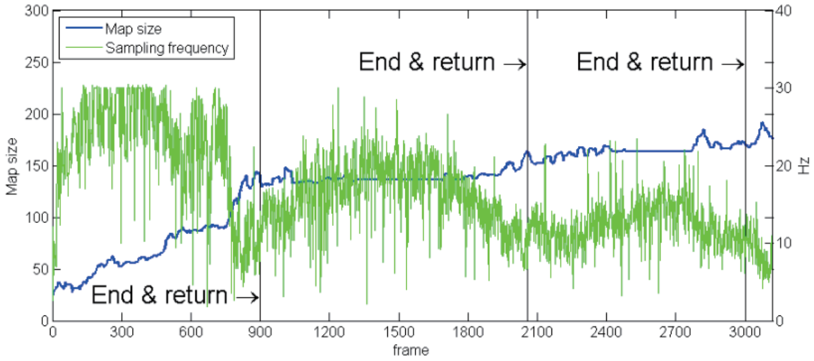


Fig. 6. (Color online) Map size and sampling frequency vs image frame.

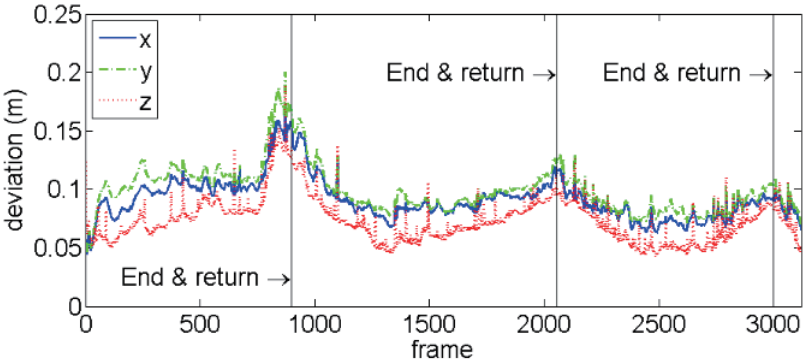


Fig. 7. (Color online) Deviations of robot's pose in *xyz*-axes.

6. Conclusions

We developed an algorithm for simultaneous localization and mapping by an aerial robot using a monocular vision sensor. In this study, we solved the problems of determining the map scale as well as initializing new landmarks by utilizing an ultrasound range detector. For the aerial robot SLAM system, the map scale was determined from the pixel coordinates of image features and the distance information provided by an ultrasonic sensing system. We also extended the usability of

SURF detectors in SLAM tasks by using its robust representation of visual landmarks. The SURF features were detected from the images to build the environmental map. For each SURF feature, the state was initialized by one 6D vector using inverse depth parameterization method. The experiments were carried out to validate the performance of the vector aerial robot SLAM systems. The experimental results showed that the EKF-SLAM can deal with the landmark initialization problem and correctly estimate the robot's pose with a standard deviation of less than 10 cm.

Acknowledgements

This work was partially supported by the Ministry of Science and Technology, Taiwan, under grant Nos. MOST 103-2221-E-032-041, MOST 103-2632-E-032-001-MY3, and MOST 104-2221-E-032-020.

References

1 A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse: IEEE Trans. Pattern Anal. **29** (2007) 1052.
2 J. Civera, A. J. Davison, and J. M. M. Montiel: IEEE T. Robot. **24** (2008) 932.
3 T. Lindeberg: Int. J. Comput. Vision **30** (1998) 79.
4 H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool: Comput. Vision Image Understanding **100** (2008) 346.
5 G. Welch and G. Bishop: An Introduction to Kalman Filter (UNC-Chapel Hill TR 95-041, Chapel Hill, North Carolina, 2006).
6 L. M. Paz, P. Pinies, J. D. Tardos, and J. Neira: IEEE Trans. Robot. **24** (2008) 946.
7 L. Sciavicco and B. Siciliano: Modeling and Control of Robot Manipulators (McGraw-Hill, New York, 1996).
8 P. A. Viola and M. J. Jones: Proc. CVPR (IEEE, 2001) p. 511.
9 G. Shakhnarovich, T. Darrell, and P. Indyk: Nearest-Neighbor Methods in Learning and Vision (The MIT Press, Cambridge, MA, 2006).