Sensors and Materials, Vol. 28, No. 9 (2016) 1013–1021
MYU Tokyo

S & M 1263

Data Association of Aerial Robot Monocular Simultaneous Localization and Mapping

Yin-Tien Wang*, Chung-Hsun Sun, Ting-Wei Chen, and Chen-Tung Chi[1]

Department of Mechanical and Electro-Mechanical Engineering, Tamkang University,
151 Ying-Chuan Rd., New Taipei City 25137, Taiwan
[1]Department of Mechanical Engineering, Taipei City University of Science and Technology,
No. 2, Xueyuan Rd., Beitou, 112 Taipei, Taiwan

(Received December 24, 2015; accepted July 6, 2016)

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

This paper presents an algorithm for data association for the visual navigation of aerial robots. The major objective is to provide the aerial robot with the capabilities of localization and mapping in global positioning system (GPS) denied environments. The visual sensor system could measure information for robot state estimation and environmental mapping as the aerial robot navigates in a GPS-denied environment. Only one single camera was used to reduce the load on the aerial robot. The captured image was transmitted to a personal computer for image processing using a radio frequency transmitter. In this study, an efficient data association method based on fuzzy rules was developed to determine the robust landmarks for robot mapping. An ultrasonic sensor was designed to provide distance measurements and to solve the map scale determination problem of monocular vision. The software program of the robot navigation system was developed on a windows-based personal computer. The navigation system integrated the visual sensor, the algorithm for data association, and the state estimator. The integrated system was used to carry out simultaneous localization and mapping for aerial robots.

1. Introduction

An aerial robot relies on sensing information to know the outside world and estimate the state of the robot itself in an unknown environment. Commonly used sensors include the global positioning system (GPS), laser range finder (LRF), and vision sensor. A GPS signal is not available for a robot in an indoor navigation environment. The LRF can offer high-precision measured data, but it is too expensive to be extensively used. The 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 was used in this study, as shown in Fig. 1, and the image was transmitted to a PC-based controller for image processing using a radio frequency module. The monocular vision sensor captured two-dimensional images but lacked depth information on the objects. Without depth information, the location of a new landmark could not be determined; furthermore, the map scale of the environment could not be initially estimated. For monocular vision, many researchers have developed time-delayed and undelayed procedures

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

ISSN 0914-4935

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

for landmark initialization.[1,2] This study used the undelayed method. The spatial coordinates of the image feature were calculated using the method of inverse depth parameterization.[2] However, the problem of determining the map scale remained unsolved. In this study, an ultrasonic sensing system was developed to provide one-dimensional distance measurements and to solve the map scale determination problem of monocular vision.

The contribution of this paper is the novel procedures for data association. To build a persistent map of an environment, an efficient procedure of data association for visual mapping was developed. The procedures of data association include a search of image features located at the predicted location in the image plane, as well as the calculation of the Euclidean distance between the descriptors of image features. Two methods based on fixed-value levels and fuzzy rules were designed for data association.

We also extended the usability of a persistent map and the data association methods developed in the tasks of simultaneous localization and mapping (SLAM). An extended Kalman filter (EKF) was used in SLAM tasks to recursively predict and estimate the robot state and the states of environmental landmarks.[3] The problem of determining the map scale as well as initializing new landmarks were also investigated for monocular vision in robot navigation.

2. Aerial Robot SLAM

Based on the sensor measurements, the states of the robot and landmarks are estimated during the SLAM tasks. In this study, a monocular vision system was used as the only measuring sensor. The monocular camera was carried by the aerial robot and modeled as a free-moving system.[1,4] The state vector $\boldsymbol{x}_k$ at time step k can be expressed as

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

where $\boldsymbol{u}_k$ and $\boldsymbol{w}_k$ are the vectors of the input and the process noise, respectively. The state vector contains the states of the robot and landmarks during the SLAM tasks:

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_{\mathrm{C}}^{\mathrm{T}} & \boldsymbol{M}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{x}_{\mathrm{C}}^{\mathrm{T}} & \boldsymbol{m}_1^{\mathrm{T}} & \boldsymbol{m}_2^{\mathrm{T}} & \cdots & \boldsymbol{m}_j^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \tag{2}$$

where $\boldsymbol{x}_{\mathrm{C}} = [\boldsymbol{r}^{\mathrm{T}}\ \boldsymbol{\phi}^{\mathrm{T}}\ \boldsymbol{v}^{\mathrm{T}}\ \boldsymbol{w}^{\mathrm{T}}]^{\mathrm{T}}$ are the robot's world coordinates, and $\boldsymbol{m}_j$ denotes the jth landmark in the

environment map $\boldsymbol{M}$. The SLAM tasks are used to estimate the state $\boldsymbol{x}_k$ of the target recursively according to the measurement $\boldsymbol{z}_k$ at k:

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

where $\boldsymbol{v}_k$ is the noise vector. Since the sensor frame was set at the camera center, the coordinates of the ith landmark in the world frame were

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

where $\boldsymbol{r}$ is the sensor's world coordinates, $\boldsymbol{R}$ is the rotational matrix,[5] and $\boldsymbol{h}_i^{\mathrm{W}}$ and $\boldsymbol{h}_i^{\mathrm{C}}$ are the ray vectors of the image features in the world and sensor frames, respectively. Since the monocular vision lacks range information, the method of initializing the visual landmarks becomes a non-trivial procedure.

In this study, an undelayed visual landmark initialization procedure based on the inverse depth parameterization was developed.[2] The 3-dimensional (3D) spatial coordinates of the feature were described by a 6D position vector

$$\hat{\boldsymbol{m}}_i = \begin{bmatrix} \hat{r}_{ix}^{\mathrm{W}} & \hat{r}_{iy}^{\mathrm{W}} & \hat{r}_{iz}^{\mathrm{W}} & \hat{\theta}_i^{\mathrm{W}} & \hat{\psi}_i^{\mathrm{W}} & \hat{\rho}_i \end{bmatrix}^{\mathrm{T}}, \tag{5}$$

where $i = 1, \ldots, n$; $\hat{\boldsymbol{r}}^{\mathrm{W}} = [\hat{r}_{ix}^{\mathrm{W}} \;\; \hat{r}_{iy}^{\mathrm{W}} \;\; \hat{r}_{iz}^{\mathrm{W}}]^{\mathrm{T}}$ represents the estimated coordinates of the camera when the feature was observed, $\hat{\rho}_i$ is the image depth of the feature, and $\hat{\theta}_i^{\mathrm{W}}$ and $\hat{\psi}_i^{\mathrm{W}}$ are the longitude and latitude angles, respectively, of the spherical coordinate system located at the camera center.

For determining the map scale in a monocular SLAM problem, we developed a one-dimensional distance detector based on ultrasound technology. The distance detector consisted of an ultrasound sensor chip, a radio frequency transmitter, and a microchip. When the aerial robot took off, the ultrasound sensor was designed to measure the distance from the ground. The SLAM task began to work when the height of the quadrotor was 1.5 m above the ground. Some image features obtained from the first image were chosen as landmarks and their states were initialized according to Eq. (4). In the equation, the depth information for image features was obtained from the ultrasound sensor. With these initial image features, the map scale was also calculated.

3. Vision-Based Mapping

Robot visual mapping needs a robust method to represent visual landmarks which are detected in images. In this study, we used the method of speeded-up robust features (SURF)[6] to find the visual landmarks for robot mapping during SLAM tasks. After the features were detected, a high-dimensional vector was computed to represent the description of the features.

To match the high-dimensional description vector of a landmark with that of an image feature, we developed the procedures of data association based on fixed-value levels and fuzzy rules. The procedures of data association include searching for image features located at a predicted location in an image plane, as well as calculating the Euclidean distance between their descriptors using the nearest-neighbor search method.[7] The matching criterion for a landmark with an image feature was defined as: the feature must be located at the predicted position, and its Euclidean distance must be within the threshold value.

3.1 Level-shifted data associtation

The data association based on fixed-value levels was designed as listed in Table 1. The concept was to design a window located at the predicted position for searching the image feature and to set a threshold value for the Euclidean distance between the descriptors, as shown in Fig. 2. Four levels were designed, as shown in Table 1. For each level, the size of the search window was increased by 10 pixels, while the threshold of the Euclidean distance was decreased by 0.03. During the data association, the first level with the window size 19 × 19 and distance threshold 0.2 was initially applied. For example, as shown in the left panel of Fig. 3, landmarks nos. 0 and 3 were successfully matched with the corresponding image features. The camera speed and acceleration were 0.33 m/s and 0.57 m/s^2, respectively. However, as shown in the right panel of Fig. 3, landmark no. 3 could not be matched with the corresponding feature when the camera speed and acceleration were

Table 1
Fixed-value levels for data association.

Levels	1st	2nd	3rd	4th
Window-size*	19 × 19	29 × 29	39 × 39	9 × 49
Threshold of Euclidean distance	0.2	0.17	0.14	0.11

*Units in pixels.

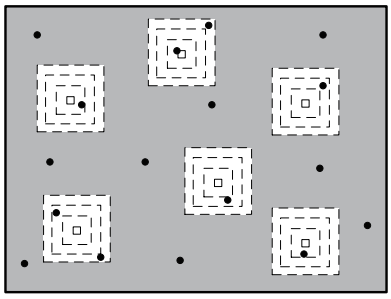
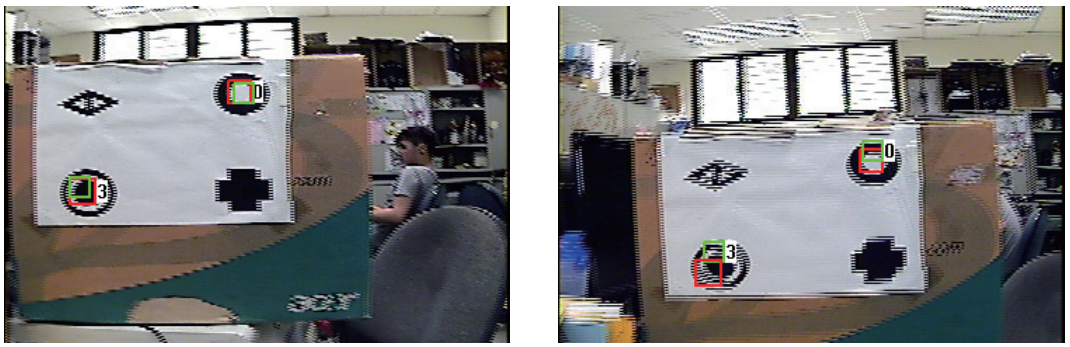


Fig. 2. Search windows for locating image features.



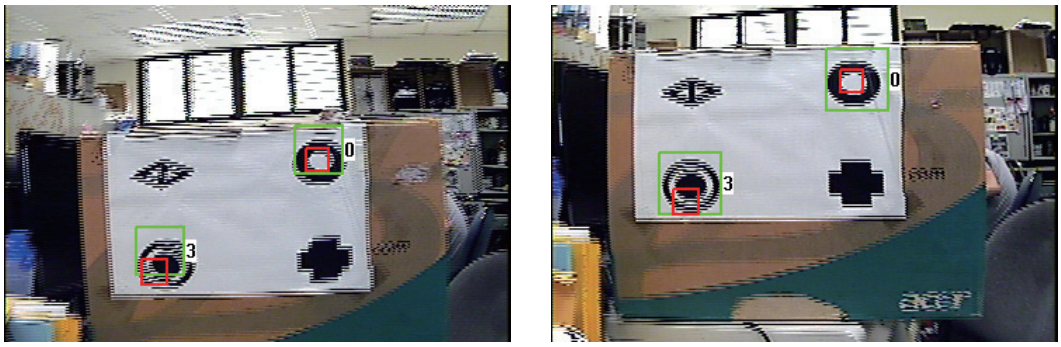
(a) (b)

Fig. 3. (Color online) (a) Camera speed of 0.33 m/s and acceleration of 0.57 m/s^2. (b) Camera speed of 0.41 m/s and acceleration of 2.14 m/s^2. In both cases, the first level was applied.

increased to be 0.41 m/s and 2.14 m/s^2, respectively. If the third level with the window size 39 × 39 and distance threshold 0.14 was applied, both landmarks nos. 0 and 3 were again matched with the corresponding features, as shown in the left panel of Fig. 4. For a higher camera speed of 0.83 m/s and an acceleration of 4.73 m/s^2, the fourth level with the window size 49 × 49 and distance threshold 0.11 had to be applied to match the corresponding features, as shown in the right panel of Fig. 4.

3.2 Fuzzy data associtation

In the level-shifted data association method, the first level had to be initially applied. If the image features were not matched successfully, then the window-size and distance threshold were shifted to the next higher levels. Therefore, the data association could not respond quickly. The data association method based on fuzzy rules was designed to improve the response speed. The velocity v_c and acceleration a_c were chosen as the inputs to the fuzzy rules. The input and output membership functions were planned, as shown in Figs. 5 and 6, respectively. The absolute velocity v_c varied from 0 to 2 m/s, while the absolute acceleration a_c changed from 0 to 4 m/s^2. The output



(a) (b)

Fig. 4. (Color online) (a) Camera speed of 0.41 m/s and acceleration of 2.14 m/s^2. The third level was applied. (b) Camera speed of 0.83 m/s and acceleration of 4.73 m/s^2. The fourth level was applied.

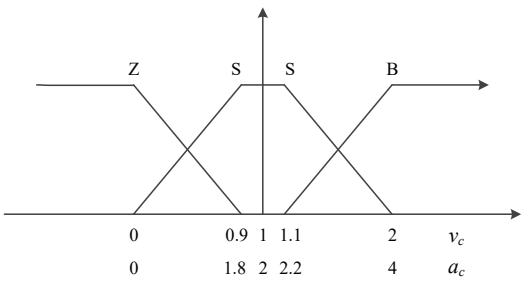


Fig. 5. Membership functions of the velocity and acceleration inputs.

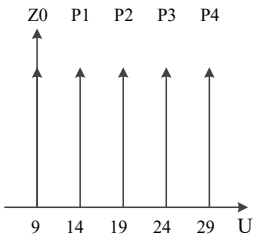


Fig. 6. Membership functions of the outputs.

U was limited from 9 to 29 pixels. The fuzzy rule base was designed according to the experiments and is listed in Table 2. The center-of-gravity method was used to defuzzify the output:

$$U = \frac{\sum_{i=1}^{n} w_i(v_c, a_c)u_i}{\sum_{i=1}^{n} w_i(v_c, a_c)}, \tag{6}$$

where w_i is the weight value of the output membership function u_i. The output U is the radius of the search window and the resultant window-size is $(2U + 1) \times (2U + 1)$ pixels. The threshold of Euclidean distance d_{match} was chosen to be

$$d_{match} = d_{match\_int} - (U - Z0)\Delta d_{match}, \tag{7}$$

where $d_{match\_int} = 0.2$ was the initial distance; $\Delta d_{match} = 0.006$ was the incremental distance, and $Z0 = 9$ was the initial value of the output membership function.

4. Results

To implement the navigating tasks, the monocular vision was integrated with the free-moving state model and the measurement model to form a SLAM system. Once the images were captured by the camera, features were detected using SURF. The system carried out data associations of the map landmarks and the image features using the proposed level-shifted and fuzzy rule methods.

Two experiments were performed to validate the proposed algorithms. The first experiment depicted the performance comparison of two developed data association methods. The aerial robot SLAM task was implemented in the second experiment to demonstrate the performance of the integrated system.

4.1 Performance of data associtation methods

The performances of two data association methods were compared in this experiment. For a scene in a SLAM task similar to that shown in Figs. 7 and 8, two data association methods were applied to locate the landmarks in the map. Using the fuzzy data association method, landmark no. 301 was identified at the lower-right corner, as shown in Fig. 7. However, the same landmark could not be identified using the level-shifted method in the first two calculations, as shown in Fig. 8.

Table 3 shows the number of features extracted using two different data association methods. To obtain sufficient robust landmarks for the environment map during the SLAM task, the level-shifted

Table 2
Table of fuzzy rule base.

(v_c)	(a_c)		
	Z	S	B
Z	Z0	P1	P2
S	P1	P2	P3
B	P2	P3	P4

Fig. 7. (Color online) Performance of feature matching by the fuzzy searching window.

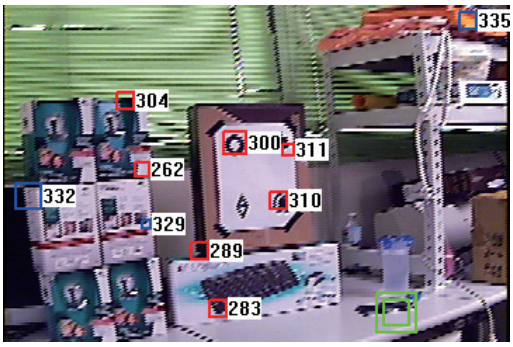


Fig. 8. (Color online) Performance of feature matching by the level-shifted researching window.

Table 3
Performance comparison of data association methods.

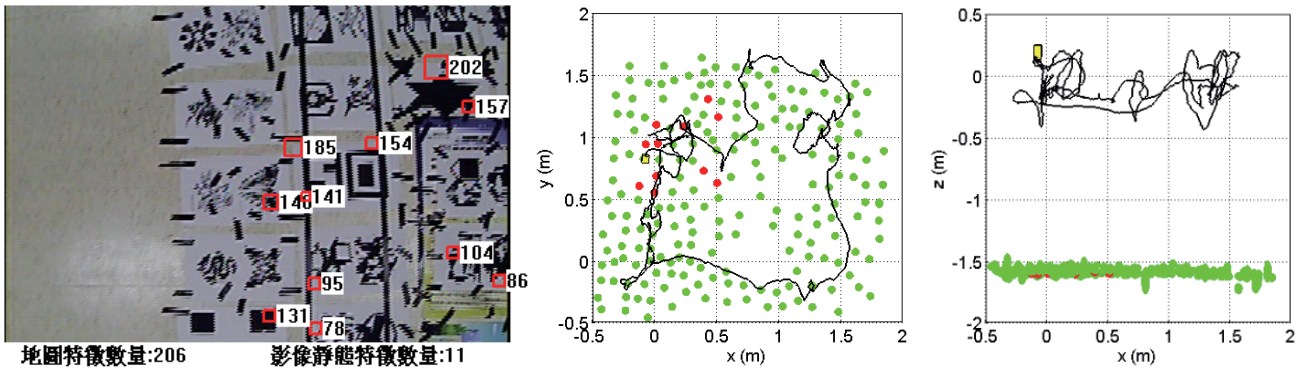
	No. of landmarks	No. of extracted features	No. of features/ No. of landmarks
1. Level-shifted	196	1119	5.71
2. Fuzzy	205	956	4.66

method had to extract 5.71 times the number of image features. On the other hand, the fuzzy method only had to extract 4.66 times the number of image features. Therefore, we concluded that the fuzzy method was more efficient than the other method in searching for robust visual landmarks.

4.2 Aerial robot SLAM

The SLAM experiment was implemented on a real quadrotor aerial robot. The monocular vision was carried by the quadrotor to follow a trajectory above a square plane with a 2 m side, as shown in Fig. 9. The camera lens always faced downward as the aerial robot flew along the planar trajectory. The SLAM system started when the quadrotor's height was 1.5 m, as measured by the ultrasound sensor. Fifteen SURF features obtained from the first image were chosen as 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 constantly added to the map and their state vectors were initialized using Eq. (5). As the monocular vision followed the planar trajectory, the SLAM system built the environmental map concurrently and estimated the robot pose. Figure 10 presents the 1650th RGB image frame obtained in the experiments. Eleven landmarks were detected in this frame, and the map size was increased to 206. The top-view (xy-plane) and side-view (yz-plane) plots of the environmental map are depicted in the middle and right panels, respectively. The map size and sampling frequency versus the image frame is plotted in Fig. 11. The map size was increased so that it contained about 200 landmarks at the end of the first loop and 206 landmarks at the end of the experiment. The average sampling frequency was about 30 Hz and the lowest frequency was about 15 Hz.

Fig. 9. (Color online) Planar trajectory of the quadrotor SLAM task.



(a) (b) (c)

Fig. 10. (Color online) (a) Image, (b) top-view map (*xy*-plane), and (c) side-view map (*yz*-plane) of 1650th frame.

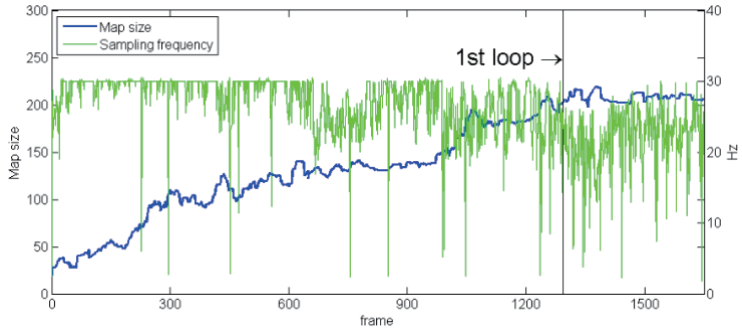


Fig. 11. (Color online) Map size and sampling frequency vs image frame.

5. Conclusions

In this study, an algorithm for data association was developed for simultaneous localization and mapping of an aerial robot with a monocular vision sensor. The algorithm was designed using fuzzy rules to construct a persistent environmental map. The fuzzy rule-based data association could efficiently search for robust visual landmarks for robot mapping within a predicted search window. We extended the usability of SURF detectors by using its robust representation of visual landmarks. For each SURF feature, the state was initialized by one 6D vector using an inverse depth parameterization method. We solved the problems of determining the map scale as well as initializing new landmarks by utilizing an ultrasound range detector. Two experiments were performed to validate the performance of the vector aerial robot SLAM systems. The experimental results showed that the data association problem could be solved and the EKF-SLAM could correctly estimate the robot's pose.

Acknowledgements

This work was partially supported by the Ministry of Science and Technology, Taiwan, under grant nos. 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 G. Welch and G. Bishop: An Introduction to Kalman Filter (UNC-Chapel Hill TR 95-041, Chapel Hill, North Carolina, 2006).
4 L. M. Paz, P. Pinies, J. D. Tardos, and J. Neira: IEEE Trans. Robot. **24** (2008) 946.
5 L. Sciavicco and B. Siciliano: Modeling and Control of Robot Manipulators (McGraw-Hill, New York, 1996).
6 H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool: Comput. Vision Image Understanding **100** (2008) 346.
7 G. Shakhnarovich, T. Darrell, and P. Indyk: Nearest-Neighbor Methods in Learning and Vision (The MIT Press, Cambridge, MA, 2006).