Proceedings of the 2nd World Congress on Electrical Engineering and Computer Systems and Science (EECSS'16)
Budapest, Hungary – August 16 – 17, 2016
Paper No. MVML 103
DOI: 10.11159/mvml16.103

Fuzzy Data Association of Aerial Robot Monocular SLAM

Yin-Tien Wang, Ting-Wei Chen
Tamkang University
151 Ying-Chuan Rd., Tamsui, New Taipei City, TAIWAN
ytwang@mail.tku.edu.tw; akhibara_moe@hotmail.com

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

Keywords**:** Visual localization and mapping; Aerial robot navigation; Detection of image features; Robot vision.

1. Introduction

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 the aerial robot, a single camera is used in this study, as shown in Fig. 1, and the image is transmitted to PC-based controller for image processing using a radio frequency module. The monocular vision sensor captures only two-dimensional images and lacks for the depth information of environmental objects. Without the depth information, the location of a new landmark cannot be determined, meanwhile the map scale of the environment cannot be estimated initially. For the monocular vision, many researchers have developed landmark initialization procedures either in time-delayed method [1] or un-delayed method [2]. The un-delayed method will be 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 is still unsolved. In this study, an ultrasonic sensing system is developed to provide one-dimensional distance measurement and solve the map scale determination problem of monocular vision.

The image features detected from the vision sensor can be used to represent the landmarks in the environment and build an environmental map for 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. The 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. [4] replaced the Gaussian second-order derivative with the box filter and calculated the approximation of determinant of the Hessian matrix using the integral image method. This method, called speeded-up robust features (SURFs), significantly reduces the calculation time. The SURF algorithm is employed in this study to detect the features from monocular RGB images and to represent the landmarks in the environmental map.

The contribution of this paper is the novel procedures of data association. In order to build a persistent environment map, an efficient procedure of data association for the SURF-based mapping is developed. The procedures of data association include the search of the image feature located at the predicted location in image plane, as well as the calculation of the Euclidian distance between SURF descriptors. Two methods based on fixed-value levels and fuzzy rules are designed for data association. Meanwhile, we also extend the usability of persistent map and the developed data association methods in the tasks of simultaneous localization and mapping (SLAM). In the SLAM tasks, the extended Kalman filter (EKF) [5] is used to recursively predict and estimate the robot state as well as the states of environmental landmarks. The problem of determining the map scale as well as initializing new landmark are also investigated for monocular vision in robot navigation.

Fig. 1: Quadrotor aerial robot with a monocular vision sensor.

2. Aerial Robot SLAM

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

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

where $\mathbf{x}_k$ is the state vector, $\mathbf{u}_k$ 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,

$$\mathbf{x} = [\mathbf{x}_C^T, \boldsymbol{M}^T]^T = [\mathbf{x}_C^T, \mathbf{m}_1^T, \mathbf{m}_2^T, \cdots, \mathbf{m}_j^T]^T \tag{2}$$

where $\mathbf{x}_C=[\boldsymbol{r}^T, \boldsymbol{\phi}^T, \mathbf{v}^T, \boldsymbol{\omega}^T]^T$ denotes the robot coordinates in world frame, and $\mathbf{m}_j$ represents the *j*th landmark in the environment map $\boldsymbol{M}$. The objective of the robot SLAM tasks is to estimate the state $\mathbf{x}_k$ of the target recursively according to the measurement $\mathbf{z}_k$ at *k*,

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

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

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

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; $\boldsymbol{h}_i^W$ and $\boldsymbol{h}_i^C$ are the ray vectors of the image features 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 becomes an important topic. In this study, a visual landmark initialization procedure based on the inverse depth parameterization [2] is developed and described in the following section.

3. Vision-Based Mapping

For the initialization of new landmarks in the monocular vision system, the un-delayed method is used 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]. We also developed a one-dimensional distance detector based on the ultrasound technology to determine the map scale in monocular SLAM problem [8]. The distance detector consists of an ultrasound sensor chip (HC-SR04), a radio frequency transmitter (3Dr Telemetry), and a microchip (Arduino Nano). 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 from the ground. At the beginning of the SLAM task, some SURF features obtained from the first image are chosen as the map landmarks and their states are initialized according to eq. (4). In the equation, the depth information of these SURF features is obtained from the ultrasound sensor. With these initial SURF features, the map scale is also calculated. After the map scale is obtained, the ultrasound sensor is turned off and the further added landmarks are initialized by using the inverse depth parameterization [2].

Robot visual mapping needs a robust method to represent the visual landmarks which are detected from images. In this study, we used the SURF method to detect and represent the 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 is further combined with the integral image method to reduce the image processing time [9]. After the features are detected from the image, the description vector is computed to represent feature characteristics. A high-dimensional description vector is used to describe the uniqueness of the feature.

For matching the high-dimensional description vectors for a pair of map landmark and image feature, this study developed the procedures of data association based on fixed-value levels and fuzzy rules, respectively. The procedures of data association include the search of the image feature located at the predicted location in image plane, as well as the calculation of the Euclidian distance between their descriptors using the nearest-neighbor search method [10]. The matching criterions for a pair of map landmark and image feature is defined as: the feature must locate at the predicted position and their Euclidian distance be within the threshold value.

3.1. Level-shifted Data Assocstation

The data association based on fixed-value levels is designed as shown in Table 1. The concepts are to design a window located at the predicted position for searching the image feature and to set a threshold value for the Euclidian distance between the descriptors. Four levels are included in Table 1. For each level, the size of the search window is increased by 10 pixels, as shown in Fig. 2, meanwhile the threshold of Euclidian distance is decreased by 0.03. During the data association, the first level with the window size 19×19 and distance threshold 0.2 is initially applied. For example, as shown in left panel of Fig. 3, landmarks no. 0 and 3 are successfully matched with the corresponding image features. The camera speed and acceleration are 0.33m/sec and 0.57 m/sec^2, respectively. However, as shown in right panel of Fig. 3, landmark no. 3 could not be matched with the corresponding features when the camera speed and acceleration are increased to be 0.41m/sec and 2.14 m/sec^2, respectively. If the third level with the window size 39×39 and distance threshold 0.14 is applied, both landmarks no. 0 and 3 are again matched with the corresponding features, as shown in left panel of Fig. 4. For higher camera speed at 0.83m/sec and acceleration at 4.73 m/sec^2, the fourth level with the window size 49×49 and distance threshold 0.11 must be applied in order to match the corresponding features, as shown in right panel of Fig. 4.

Table 1: Fixed-value levels for data association.

Levels	1st	2nd	3rd	4th
Window-size*	19×19	29×29	39×39	49×49
Thresholdof Euclidiandistance	0.2	0.17	0.14	0.11

* Unit in pixels.

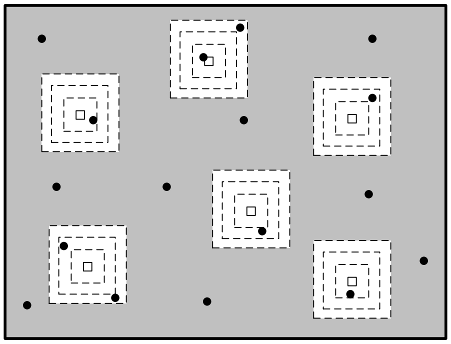

Fig. 2: Search windows for locating the image features.

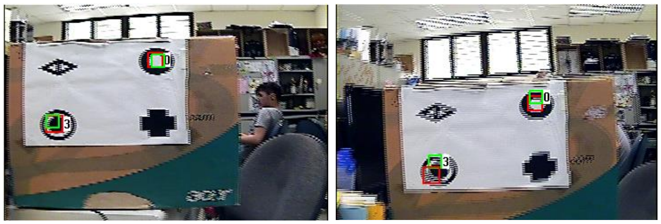

Fig. 3: Left panel: camera speed at 0.33m/sec and acceleration at 0.57 m/sec^2; Right panel: camera speed at 0.41m/sec and acceleration at 2.14 m/sec^2. In both cases, the first level is applied.

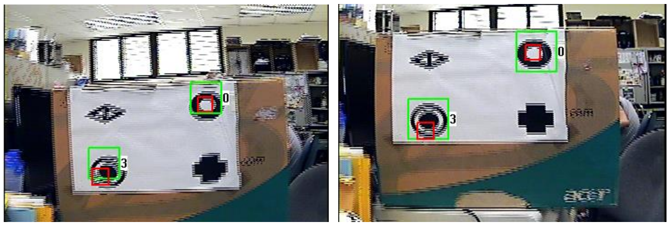

Fig. 4: Left panel: camera speed at 0.41m/sec and acceleration at 2.14 m/sec^2. The third level is applied; Right panel: camera speed at 0.83m/sec and acceleration at 4.73 m/sec^2. The fourth level is applied.

3.2. Fuzzy Data Associtation

In the level-shifted data association method, the first level must be initially applied. If the image features are not matched successfully, then the window-size and distance threshold are shifted to higher levels. Therefore, the data association cannot response quickly. The data association method based on fuzzy rules is designed to improve the response speed. The velocity v_c and acceleration a_c are chosen as the inputs of the fuzzy rules. The input and output membership functions are planned as shown in Figs.5 and 6, respectively. The absolute velocity v_c varies from 0 to 2m/sec, while the absolute acceleration a_c changes from 0 to 4m/sec^2. The output U is limited from 9 to 29 pixels. The fuzzy rule base is designed according to the experiments and listed in Table 2. The center-of-gravity method is used to defuzzify the output,

$$U = \frac{\Sigma_{i=1}^{n} w_i(v_c, a_c) u_i}{\Sigma_{i=1}^{n} w_i(v_c, a_c)} \tag{5}$$

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 Euclidian distance d_{match} is chosen to be

$$d_{match} = d_{match\_int} - (U - \mathrm{Z0})\Delta d_{match} \tag{6}$$

where d_{match_int}=0.2 is the initial distance; Δd_{match}=0.006 is the incremental distance; Z0=9 is the initial value of output membership function.

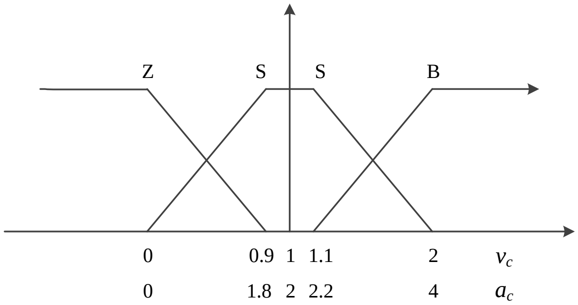

Fig. 5: Membership functions of the velocity and acceleration inputs.

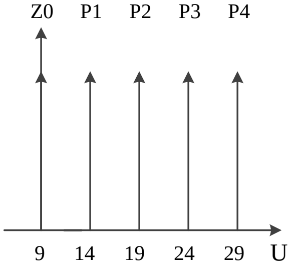

Fig. 6: Membership functions of the outputs.

Table 2: Table of fuzzy rule base.

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

4. Experimental Results

For implementing the navigating tasks, the monocular vision is integrated with the free-moving motion model, the measurement model, and the SURF detection algorithm to form a SLAM system. Once the images are captured by the camera, image features are detected by using the SURF method. The system performs data association of the map landmarks and the image features using the proposed level-shifted and fuzzy rule methods. A map management system is also designed to coordinate the newly added features and the "bad" features in the system. New features are chosen as landmarks and added to the map when the robot explores an unknown environment. The state variables of all new landmarks are augmented in the state vector in eq. (1). However, features that are not continuously detected during the task are considered as "bad" features and are deleted from the state vector.

Two experiments are carried out to validate the proposed algorithms. The first experiment depicts the performance comparison of two developed data association methods. Aerial robot SLAM task is implemented in the second experiment to demonstrate the performance of the integrated system.

4.1. Performance of Data Associtation Methods

The performances of two developed data association methods are compared in this experiment. For the similar scene in a SLAM task as shown in Figs. 7 and 8, two data association methods are performed to locate the landmarks in the map. By using the fuzzy data association method, landmark no. 301 is determined at the down-right corner, as shown in Fig. 7.

However, the same landmark could not be obtained by using the level-shifted method at the first two calculations, as shown in Fig. 8.

Table 3 depicts the performance comparison of number of features extracted by using two different data association methods. In order to obtain enough robust landmarks for the environment map during a SLAM task, the level-shifted method has to extract 5.71 times of the number of image features. On the other hand, the fuzzy method only has to extract 4.66 times of the number of image features. Therefore, it is concluded that the fuzzy method is more efficient than the other in searching for the robust visual landmarks.

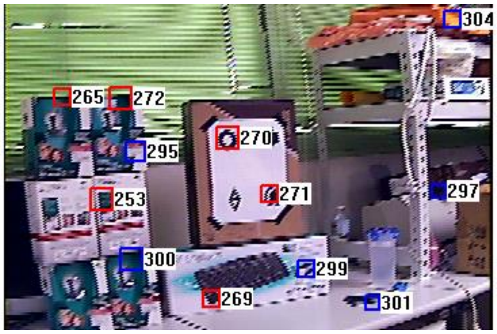

Fig. 7: Performance of feature matching for fuzzy searching window.

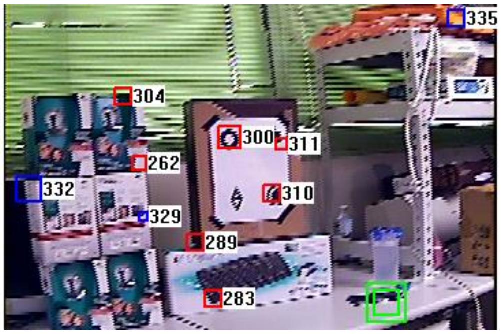

Fig. 8: Performance of feature matching for level-shifted researching window.

Table 3: Performance comparison of data association methods.

	No. of landmarks	No. of extracted features	$\frac{\text{No.of features}}{\text{No.of landmarks}}$
1. Level-shifted	196	1,119	5.71
2. Fuzzy	205	956	4.66

5. Conclusions

We developed an algorithm for aerial robot simultaneous localization and mapping using a monocular vision sensor. In this paper, we developed the procedures of data association to construct a persistent environment map. The fuzzy rule-based data association could efficiently search for the robust visual landmarks for robot mapping within a predicted search window. We also extend 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. 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. Two 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 data association problem and correctly estimate the robot 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, “MonoSLAM: Real-Time Single Camera SLAM,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1052-1067, 2007.

[2] J. Civera, A. J. Davison and J. M. M. Montiel, “Inverse Depth Parametrization for Monocular SLAM,” *IEEE Transactions on Robotics*, vol. 24, pp. 932-945, 2008.

[3] T. Lindeberg, “Feature detection with automatic scale selection,” *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79-116, 1998.

[4] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, “Speeded-Up Robust Features (SURF),” *Computer Vision and Image Understanding*, vol. 110, pp. 346-359, 2008.

[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, “Large-Scale 6-DOF SLAM with Stereo-in-Hand,” *IEEE Transactions on Robotics*, vol. 24, pp. 946-957, 2008.

[7] L. Sciavicco and B. Siciliano, *Modeling and Control of Robot Manipulators*, New York: McGraw-Hill, 1996.

[8] Y. T. Wang, C. H. Sun and T. W. Chen, “Determination of Map Scale and Initialization of Landmarks for Aerial Robot Monocular Visual Localization and Mapping,” *Sensors and Materials*, vol. 28, no. 6, pp. 667-674, 2016.

[9] P. A. Viola and M. J. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511-518.

[10] G. Shakhnarovich, T. Darrell and P. Indyk, *Nearest-Neighbor Methods in Learning and Vision*, Cambridge, MA: MIT Press, 2006.