Robotica (2014) volume 32, pp. 533–549. © Cambridge University Press 2013
doi:10.1017/S0263574713000830

Improvement of speeded-up robust features for robot visual simultaneous localization and mapping

Yin-Tien Wang* and Guan-Yu Lin

Department of Mechanical and Electro-Mechanical Engineering, Tamkang University, New Taipei City 25137, Taiwan

(Accepted July 31, 2013. First published online: September 2, 2013)

SUMMARY

A robot mapping procedure using a modified speeded-up robust feature (SURF) is proposed for building persistent maps with visual landmarks in robot simultaneous localization and mapping (SLAM). SURFs are scale-invariant features that automatically recover the scale and orientation of image features in different scenes. However, the SURF method is not originally designed for applications in dynamic environments. The repeatability of the detected SURFs will be reduced owing to the dynamic effect. This study investigated and modified SURF algorithms to improve robustness in representing visual landmarks in robot SLAM systems. Many modifications of the SURF algorithms are proposed in this study including the orientation representation of features, the vector dimension of feature description, and the number of detected features in an image. The concept of sparse representation is also used to describe the environmental map and to reduce the computational complexity when using extended Kalman filter (EKF) for state estimation. Effective procedures of data association and map management for SURFs in SLAM are also designed to improve accuracy in robot state estimation. Experimental works were performed on an actual system with binocular vision sensors to validate the feasibility and effectiveness of the proposed algorithms. The experimental examples include the evaluation of state estimation using EKF SLAM and the implementation of indoor SLAM. In the experiments, the performance of the modified SURF algorithms was compared with the original SURF algorithms. The experimental results confirm that the modified SURF provides better repeatability and better robustness for representing the landmarks in visual SLAM systems.

KEYWORDS: Robot mapping; Speeded-up robust features (SURF); Simultaneous localization and mapping (SLAM); Image processing; Robot vision.

1. Introduction

Building a persistent map with robust image features (landmarks) is an important step when implementing a visual simultaneous localization and mapping (SLAM) system. The image features require high repeatability and unique descriptions to ensure successful detection at each time step. This study developed algorithms for building a persistent map for robotic visual SLAM system with local invariant feature detectors. To perform SLAM tasks, the environmental characteristics are represented using a sparse but persistent map for real-time implementation. The advantage of spare representation is that the computational cost for information update could be reduced when using extended Kalman filter (EKF) in SLAM systems.

Many detection methods have been used for visual landmark detection in robot SLAM. For visual SLAM tasks,[4,11] the most popular of these methods is the Harris method[6] of detecting image corners by calculating the eigenvalues of the second moment matrix. Although the Harris corner detector is a simple and easily implemented algorithm for performing robot SLAM tasks, its scale-variant method of corner detection incurs a high computational cost of recovering the scale and orientation of corner features in SLAM tasks. For example, it has limited robustness for tracking features when the camera

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

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

is moving because, as the distance and angle of camera viewpoint change, the scale and orientation of the corner feature change, which can cause an image feature tracking failure. Therefore, improved methods are needed to recover the scale and orientation of image features.[4] In contrast, a detection method using local invariant features[8,9] automatically resolves these problems and provides a robust representation method for scale- and orientation-invariant features in SLAM task.[7]

Lindeberg proposed the concept of automatic scale selection to overcome the disadvantage of Harris corner[8]. By using the convolution of the image and Laplacian of Gaussians (LoG) to establish a Hessian matrix, the local invariant features were detected by calculating the determinant of the Hessian matrix. Although the Lindeberg method provides stability and high repeatability, its shortcoming is the computational inefficiency of the local invariant feature detectors. Several approaches were later proposed to improve the processing speed. For example, Lowe[9] replaced LoG by Difference of Gaussians (DoG) to reduce image convolution complexity; to improve the computational performance, Bay *et al.*[2] applied the concepts of box filter and integral image[13] in their speeded-up robust features (SURF) approach for detecting local invariant features. Although the literature suggests that SURF is superior to other methods of detecting and representing image point features,[5] the computational cost of the SURF method in visual SLAM has not been evaluated. Many studies that have applied local invariant features in robot localization and mapping tasks include Karlsson *et al.*,[7] who developed a visual SLAM system that uses a scale-invariant feature transform (SIFT) method of detecting local invariant features, which was developed earlier by Lowe.[9] Murillo *et al.*[10] used the SURF method instead of the SIFT method for image feature detection and recognition. Their experimental results showed improved efficiency in robot visual localization. Wu and Fu[14] used SURF to build the frontal system for visual SLAM and to provide stable static landmarks. Artiera *et al.*[1] used feature clusters containing both Harris and SURF descriptors to represent the map of different environments and implemented visual SLAM from a mini UAV. Cocaud and Kubota[3] developed a SURF-based SLAM scheme for autonomous landing on asteroids. However, real-time implementation and performance evaluation of SURF-based SLAM were not performed in a dynamic environment.

This study proposes a novel algorithm for using the modified SURF algorithms for map building. The feature detection, description, and recognition procedures are evaluated and modified to optimize robustness in representing visual landmarks. The modifications to the SURF algorithms include the orientation representation of features, the vector dimension of feature description, and the number of detected features in an image. The proposed algorithms are also validated in experiments performed on an actual system. Notably, all visual SLAM tasks in the experiments are performed using only a binocular vision system as the sensing device.

The contribution of this paper is the novel use of modified SURF algorithms to solve the problem of recovering the scale and orientation of image features. This study also extends the usability of local invariant feature detectors in SLAM tasks by utilizing its robust representation of visual landmarks. The robustness of SLAM systems is also improved by using the proposed data association and map management for SURF-based mapping. We believe that the findings of this study are an important contribution to the literature on visual SLAM, navigation systems, and, most importantly, applications of SLAM in dynamic environments.

The remaining sections of the paper are organized as follows: the conventional SURF method and the proposed modifications of the SURF method are described in Sections 2 and 3. Section 4 presents the structure of the developed robot SLAM system. The results of three experimental applications of the proposed algorithms are then presented in Section 5. Finally, concluding remarks are given in the last section.

2. Speeded-Up Robust Features

The conventional SURF method developed by Bay *et al.*[2] uses a box filter instead of DoG to approximate the determinant of the Hessian matrix. Feature detection is performed at the pixel location where the determinant of the Hessian matrix is a local maximum value. The box filter is further combined with integral image method[13] to reduce the image processing time. Bay *et al.* applied the octaves concept to the design of box filter sizes for detecting features in different scales. After the features are detected from the image, the description vector is computed to represent feature characteristics. Bay *et al.*[2] described the orientation of a feature by using Haar wavelet filter to

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press



Fig. 1. (Colour online) Feature orientation changes caused by motion-induced blur.

compute the wavelet responses of the image feature. The orientation of a feature is defined as the direction with the largest sum of the Haar wavelet responses. A high-dimensional description vector is then used to describe the uniqueness of the feature. A square area is chosen with its center located in the feature point and its coordinates aligned with the orientation of the feature. This square area is then divided into 4×4 sub-areas. For each sub-area, a 4-dimensional descriptor vector is assigned and expressed as $v = (\sum \mathrm{dx}, \sum \mathrm{dy}, \sum |\mathrm{dx}|, \sum |\mathrm{dy}|)$, where $\sum \mathrm{dx}$, $\sum \mathrm{dy}$, $\sum |\mathrm{dx}|$, and $\sum |\mathrm{dy}|$ are the sums of the Haar wavelet responses and their absolute values in the x and y directions, respectively. All responses are defined in the coordinate system aligned with the orientation of the feature, which results in dimensions of $4 \times 4 \times 4 = 64$ for the description vector of an image feature. For matching high-dimensional description vectors, the most popular method is nearest-neighbor (NN) search method.[12] For a space of n dimensions with point sets P and Q, the query point q belongs to Q. The NN search method can be used to find the point p in P, which has minimum distance from q. The distance between the arbitrary point p in the P set and the query point q is usually defined by the norm l_s. In the case of l_2, the distance becomes Euclidean distance d. The criterion for matching two image features is usually to determine the smallest distance between their descriptors.

Mapping in visual SLAM requires a robust method of representing visual landmarks detected in an image. This study used the SURF method to detect and represent visual landmarks of a map during SLAM tasks. However, experimental works show that the conventional SURF method developed by Bay *et al.*[2] does not inherently meet mapping requirements for visual SLAM. For example, Fig. 1 shows how camera motion can cause image blurring in SLAM. The images were captured with a CMOS webcam with a 320×240-pixel resolution. The left and right panels of Fig. 1 are the clear and blurred images, respectively, of the same scene. After applying the SURF method to detect image features, however, fewer image features are detected in the blurred image compared to the clear image. The red (dark) circles in the figure represent the detected SURFs and the direction of the green (gray) radius indicates the orientation of each feature. The figure also shows that the orientation of feature numbers 20, 23, and 27 in the left panel differs from those in the right panel. If the orientations of two corresponding features in the left and right panels differ, their description vectors differ. Therefore, these two corresponding features cannot be correctly matched during the data association process.

3. Modified Speeded-Up Robust Features

In this research, the conventional SURF method was modified to improve the robustness of feature representation in SLAM tasks. The modified SURF method is described in the subsections below.

3.1. Robust orientation representation

The magnitude of total Haar wavelet response in the square area of a SURF is calculated as

$$M = (\Sigma \mathrm{d}X)^2 + (\Sigma \mathrm{d}Y)^2, \tag{1}$$

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

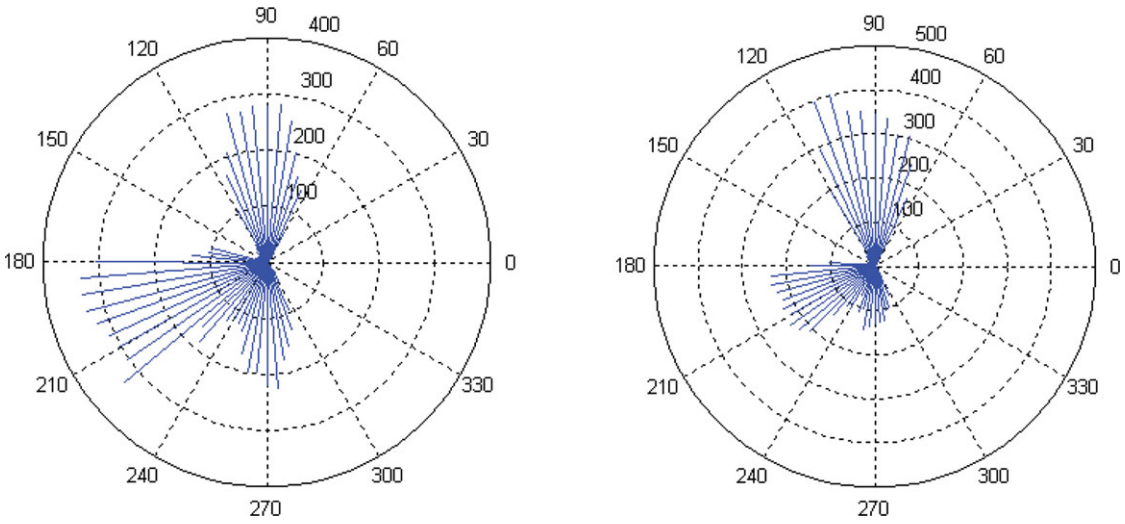


Fig. 2. (Colour online) Magnitude of total Harr wavelet response plotted in polar coordinates for feature number 27.

where dX and dY are the Haar wavelet responses in the x- and y-axes, respectively. Figure 2 plots the magnitude of wavelet response M in terms of polar coordinates. The angle of the polar coordinate represents the rotation degree of the $\pi/3$ sliding orientation window defined in ref. [2]. The left and right panels of Fig. 2 denote the wavelet response of feature number 27 in the clear and blurred images in Fig. 1, respectively. The left panel of Fig. 2 shows that, for the feature obtained from the clear image, the magnitude of wavelet response is maximum when the rotation of the sliding orientation window is $\theta_{\max} = 190°$. In contrast, for the feature obtained for the blurred image, the magnitude of the wavelet response is maximum when $\theta_{\max} = 105°$ (right panel, Fig. 2). This phenomenon of orientation change for the same feature results from motion-induced blurring in images. Bay *et al.*[2] defined the orientation of a SURF $\phi_{\max}$ located at angle $\theta_{\max}$, where the Haar wavelet response is a local maximum value as

$$\phi_{\max} = A \tan((\Sigma \mathrm{d}Y)^2, (\Sigma \mathrm{d}X)^2)\big|_{\theta_{\max}} . \tag{2}$$

The orientations of feature number 27 in Fig. 1 were calculated using Eq. (2). In this case, $\phi_{\max} = 170°$ for feature number 27 detected in the clear image and $\phi_{\max} = 270°$ for that detected in the blurred image. The different orientations for the same feature produce different descriptor vectors. This phenomenon also causes a mismatch of the same feature in two images. For example, assume that feature number 27 in the left panel of Fig. 1 was chosen as one landmark and stored in the map. Note that the sliding orientation window is at $\theta_{\max} = 190°$ and that the orientation of the feature is $\phi_{\max} = 170°$. When the same feature is detected from another image, e.g., the image in the right panel of Fig. 1, the sliding orientation window and the orientation of the feature are $\theta_{\max} = 105°$ and $\phi_{\max} = 270°$, respectively. The NN search method is used to match features by determining the smallest Euclidean distance d between two descriptor vectors. In this case, the Euclidean distance between two descriptor vectors for feature number 27 in the left and right panels of Fig. 1 was calculated as $d = 0.52$. If the Euclidean distance d between two descriptor vectors exceeds 0.3, an empirical value for this particular scene, we assume that the two features do not match because of their different appearance. Figure 3 depicts the magnitude M of wavelet response and Euclidean distance d *versus* the rotation degree θ of sliding orientation window for feature number 27. The circular symbol in the figure represents the magnitude of the wavelet response, and the diamond symbol denotes the Euclidean distance. The calculation results show that other than the rotation degree θ located in between 185° and 220° , the Euclidean distance exceeds 0.3 and the feature cannot be matched correctly.

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

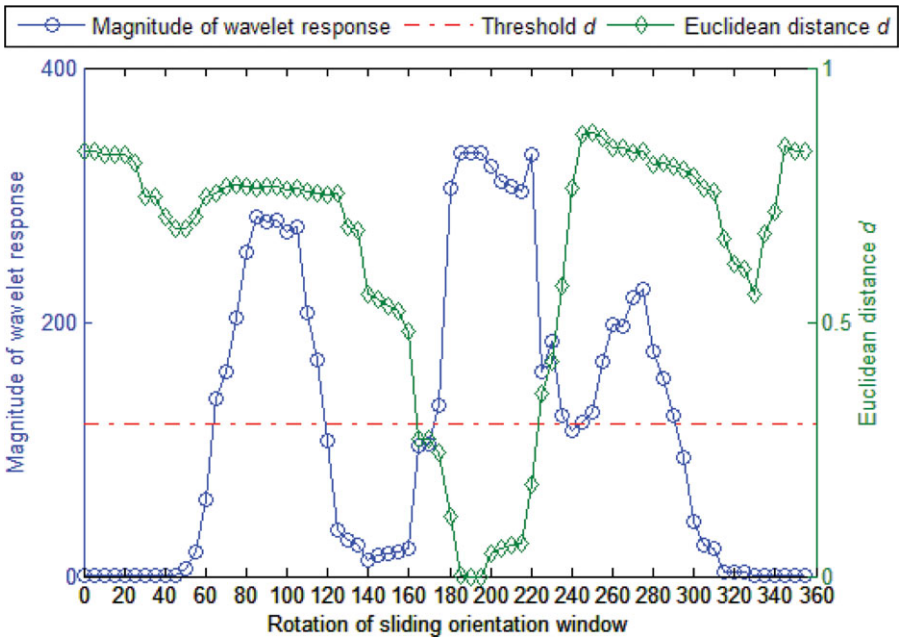


Fig. 3. (Colour online) Magnitude of wavelet response and Euclidean distance with respect to the rotation of the sliding orientation window for feature number 27.

The problem of orientation change in blurred images was solved by modifying the definition of feature orientation. The robust mean (RM) of rotation degree θ_{RM} is defined as

$$\theta_{RM} = \frac{\sum \theta_j M_j}{\sum M_j} : \left|\theta_j - \theta_{\max}\right| \leq \varepsilon, \tag{3}$$

where θ_j is the degrees of rotation in the sliding orientation window, $\theta_j = 0° - 360°$ and ε is the threshold rotation value. The corresponding feature orientation ϕ_{RM} can be determined at θ_{RM} using Eq. (2). The weighted mean (WM) of rotation degree θ_{WM} is defined as

$$\theta_{WM} = \frac{\sum \theta_j M_j}{\sum M_j}. \tag{4}$$

The corresponding feature orientation ϕ_{WM} at θ_{WM} can be determined using Eq. (2). Figure 4 plots the variation in orientations $\phi_{\max}$, ϕ_{RM}, and ϕ_{WM} for feature number 27. In this example, the camera remained horizontal and was not rotated during movement. For each scene, 250 image frames were captured at a frame rate of 10 Hz. The weighted mean orientation ϕ_{WM} shows higher stability compared to $\phi_{\max}$ and ϕ_{RM}. The feature orientation in Fig. 1 was reproduced by using weighted mean method as shown in Fig. 5. When weighted mean method is used, the orientations of feature numbers 20, 23, and 27 in the left panel are very close to those in the right panel.

3.2. Computationally efficient description vectors

The Haar wavelet response is the convolution of the Haar wavelet filter with a square area around the image feature. Computation time depends on the size of the square area. To improve computation speed, the 64-dimensional (64D) description vector was reduced to a 16-dimensional (16D) vector. Figure 6 shows the reduced square area involved in the feature area. Only the central 2×2 sub-areas are selected from the 4×4 sub-areas of a conventional SURF. Therefore, the description vector dimensions are reduced to $2 \times 2 \times 4 = 16$. Table I compares the processing times when calculating the 16D and 64D description vectors. The number of detected features varies from 100 to 800. The experimental results show that computation for the 16D description vectors is about 40%–47% faster than that for the 64D description vectors. The experiments described below also confirm that the feature with the 16D description vector has higher repeatability in the scenes used to evaluate

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

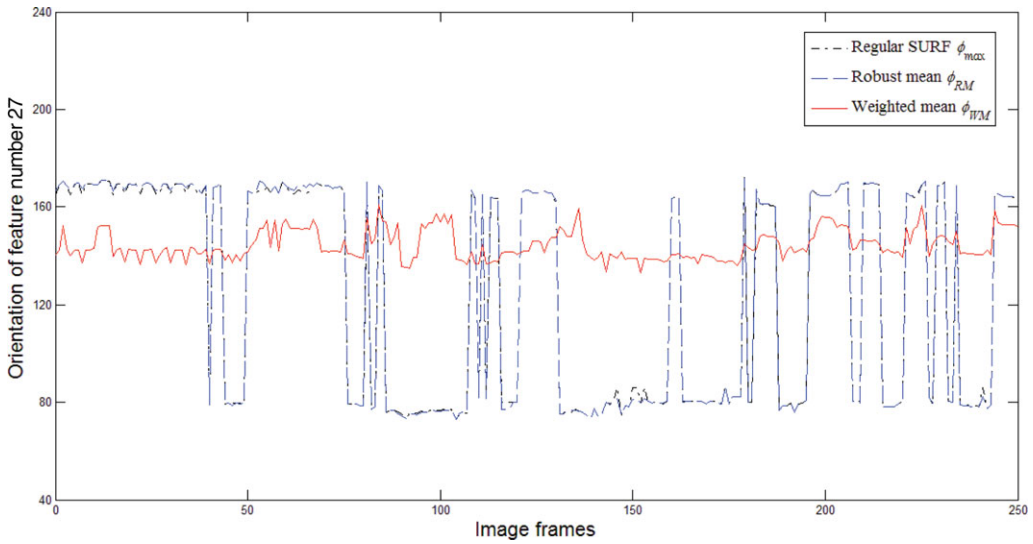


Fig. 4. (Colour online) Conventional and modified feature orientations plotted for feature number 27.



Fig. 5. (Colour online) Modified orientation of features in clear and burred images.

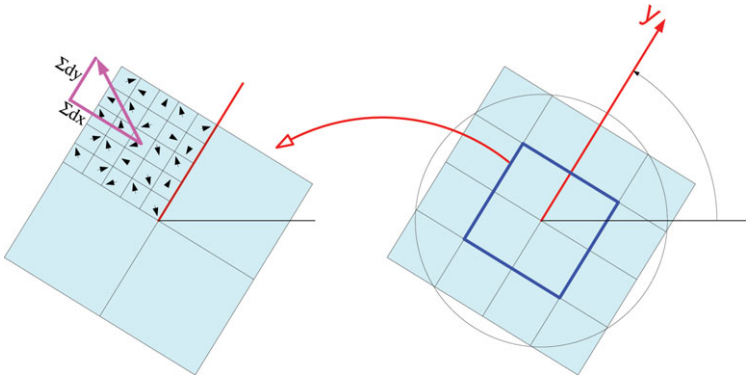


Fig. 6. (Colour online) 2 × 2 sub-areas in the center of feature area (left) and original 64D descriptor (right).

performance in SLAM tasks. Two description vectors are used in the experiments for representing each landmark to increase repeatability. The first description vector is established when the landmark is initialized. The second description vector is established after each data association process is successfully matched.

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

Table I. Comparison of processing time for 16D and 64D descriptors.

Octaves = 4			
No. of detected features	Processing time for 16D descriptors (ms)	Processing time for 64D descriptors (ms)	Time saving (%)
100	11.39	21.60	47.27
200	22.41	40.29	44.38
300	32.51	57.70	43.66
400	42.77	75.03	43.00
500	52.63	91.02	42.18
600	62.11	106.26	41.55
700	71.09	119.65	40.59
800	79.84	113.34	40.12

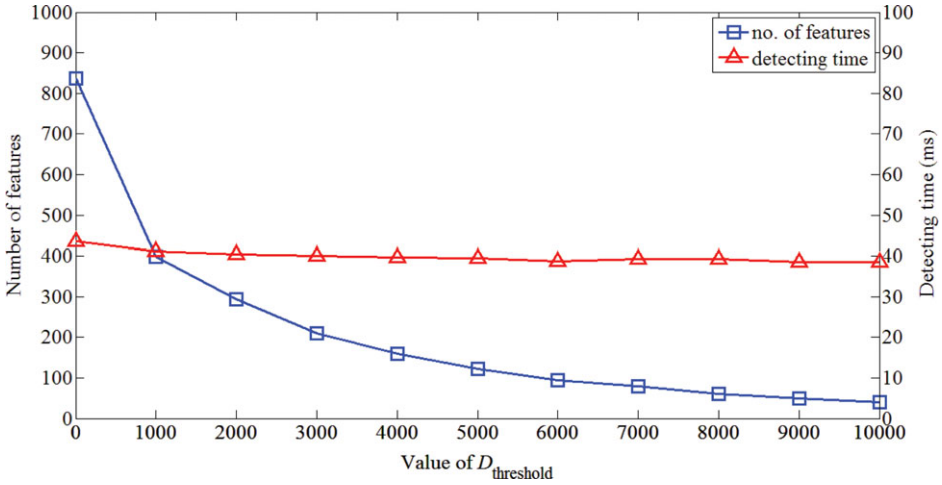


Fig. 7. (Colour online) Number of detected features and the corresponding detection time using different $D_{threshold}$ values.

3.3. Flexible feature number thresholding

For the feature detection process, a threshold value ($D_{threshold}$) was set as the lowest limit of Hessian matrix determinant for flexibly controlling the number of features detected in an image. The value of $D_{threshold}$ can be adjusted online according to environmental conditions. The objective is to determine a sufficient number of features even when the background is dull or when the environment has few features. An example of an implementation is given to demonstrate the concept. The SURFs are detected using different $D_{threshold}$ values. Figure 7 shows the number of detected features and the corresponding detection time. The detection results show that the number of detected features varied from 837 to 41 when the value of $D_{threshold}$ was changed from 0 to 10,000. Meanwhile, the computation time decreased from 43.5 ms to 38.4 ms when the number of detected features decreased. Therefore, the number of detected features can be effectively controlled, and the computational cost can be reduced by varying the $D_{threshold}$ value. Additionally, the use of box filters was limited to two low-level octaves for detecting image features. The features detected by using box filters in Octaves 3 and 4 usually have large radii. Features with a large radius are undetectable by the SURF method when they are located in the boundary region of the image. Therefore, the box filters in Octaves 3 and higher-level octaves are not used here.

4. Robot SLAM

When the robot performs SLAM tasks, the states of the robot and landmarks in the environment are estimated based on the measurement information. The state sequence of a system at time step k can

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

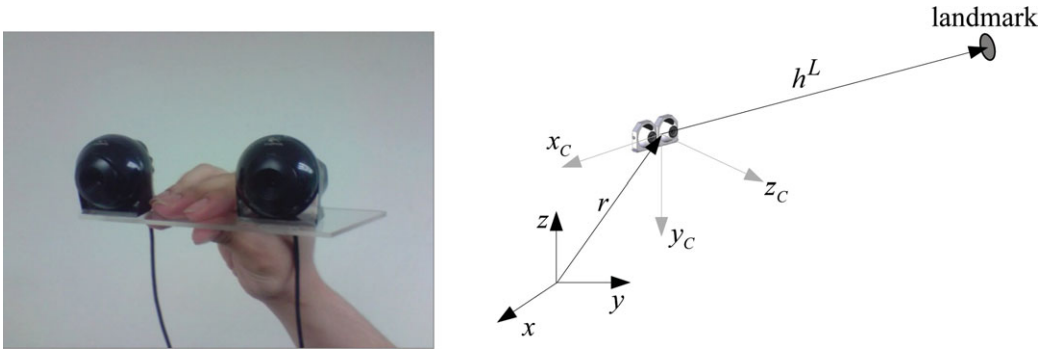


Fig. 8. (Colour online) Binocular vision sensor system.

be expressed as

$$\mathbf{x}_k = f\left(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}, w_{k-1}\right), \tag{5}$$

where $\mathbf{x}_k$ is the state vector, $\mathbf{u}_k$ is the input, and w_k is the process noise. When perform SLAM tasks using a camera, the state vector contains the states of the camera and landmarks,

$$\mathbf{x} = \left[\mathbf{x}_C^T,\ M^T\right]^T = \left[\mathbf{x}_C^T,\ \mathbf{m}_1^T,\ \mathbf{m}_2^T, \ldots, \mathbf{m}_j^T\right]^T, \tag{6}$$

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

$$\mathbf{z}_k = g\left(\mathbf{x}_k, v_k\right), \tag{7}$$

where v_k is the measurement noise. A binocular vision system (Fig. 8) is the only sensing device considered in the recursive state estimation algorithm. At time $t = k$, the vectors of measurement $\mathbf{z}_k$ and the ith observed image feature are, respectively,

$$\mathbf{z}_k = \left[\mathbf{z}_{1k}^T\, \mathbf{z}_{2k}^T \cdots \mathbf{z}_{mk}^T\right]^T; \quad \mathbf{z}_{ik} = \begin{bmatrix} I_{ix} \\ I_{iy} \end{bmatrix}, \tag{8}$$

where $i = 1, 2, \ldots, m; m$ is the number of measurements at time k, and $(I_x,\ I_y)$ are the pixel coordinates of a feature in the image plane. In this study, the hand-held binocular vision sensor is the only sensing device used for measurement in the SLAM system. The hand-held camera is treated as a free-moving robot system with unknown inputs. System states are estimated by using the EKF estimator to solve the target tracking problem.[4,11]

The implementation of the SLAM system in a binocular vision system integrates the motion model, the measurement model, and the SURF detection algorithm. Figure 9 is a flowchart of the developed SLAM system. After images are captured by the binocular camera, feature extraction is performed by the SURF method. The system performs data association of the landmarks in the map database and the image features of the extracted SURF using the proposed matching criterion. A map management system is also designed to coordinate the newly extracted features and the "bad" features in the system. After the properties of newly extracted features are investigated, a detection algorithm is used to distinguish moving and stationary objects. The state variables of all stationary landmarks are augmented in the state vector in Eq. (5). However, features that are not continuously detected at each time step are considered as "bad" features and are deleted from the state vector.

5. Experimental Results

This section describes the results of three experiments performed to validate the performance of the proposed algorithm in terms of repeatability of image features, ground truth, and robot SLAM.

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

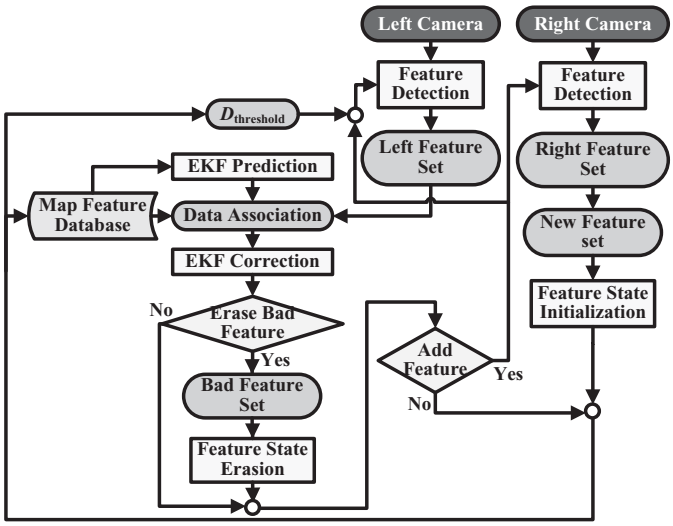


Fig. 9. Flowchart of the robot SLAM.

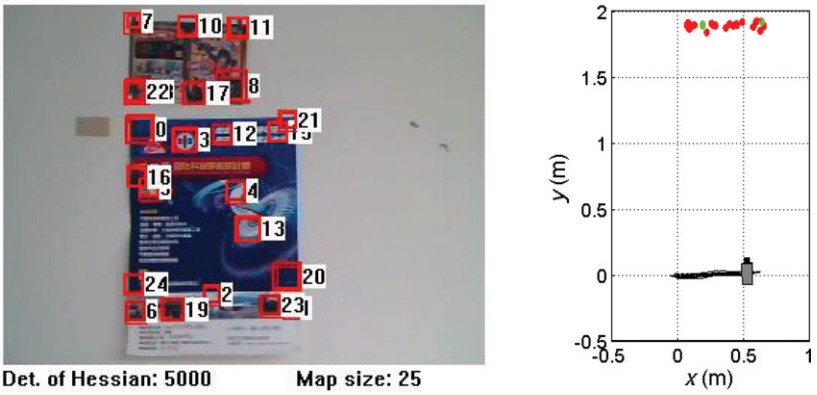


Fig. 10. (Colour online) Conventional SURF method with 64D descriptor.

5.1. Repeatability

The first experiment compared the repeatability of image features obtained by two different SURF methods, i.e., conventional SURF with 64D descriptors and modified SURF with 16D descriptors. The wall was located approximately 1.9 m away from the binocular vision system, and stationary image features were attached to the wall. Localization of the binocular vision and mapping of the image features in the environment were performed simultaneously in the experiment. The binocular vision system was hand carried from the front of the wall from point $(x, y) = (0, 0)$ to point (0.6 m, 0) and back to point (0, 0). In this experiment, 1103 images were captured in six repetitions of the motion. The maximum number of features in each image was set to 30, and the threshold of the determinant of the Hessian matrix was set to 5000. Figures 10–11 show the SLAM results for the last image frames (frame 1103) obtained by SURF and modified SURF, respectively. The panels on the right side of the figures show the estimated trajectories of the binocular vision and the feature map, and the left panels show the detected features on the image. Figure 12 compares the repeatability of the two methods. For two sequential images, some features detected from the first image were matched to features obtained from the second image. Repeatability is defined as the ratio of the number of matched features in the second image to the number of features detected in the first image. The experimental results in Fig. 12 show that the modified SURF method has average repeatability of 90%, while the conventional SURF method only has average repeatability of 80%. A low repeatability increases the number of features mapped during a SLAM task and increases the computational cost of EKF estimation. The left panel of Figs.10 and 11 shows that the map size is 25 and 19 in the conventional SURF method and the modified SURF method, respectively.

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

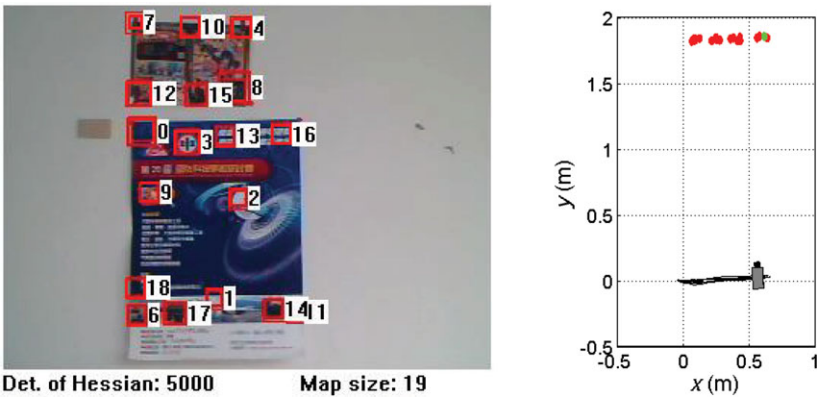


Fig. 11. (Colour online) Modified SURF with ϕ_{WM} and 16D descriptor.

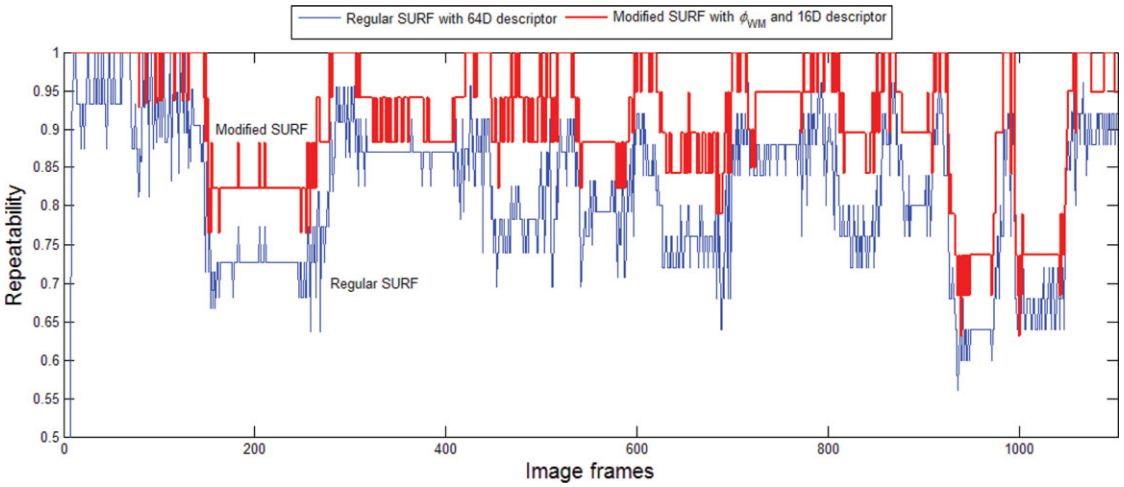


Fig. 12. (Colour online) Repeatability of conventional SURF and modified SURF *versus* image frames.

5.2. Ground truth

Figure 13 shows how the ground truth experiment was implemented on a circular path. Nine ground truth positions (points A–I) with known 3D coordinates were arranged on the circular path. The objective of this experiment is to evaluate the effectiveness of the SLAM scheme integrated with SURF detection method for estimating the states of these nine ground truth positions. The performance evaluation of the two different SURF methods integrated with SLAM system included conventional SURF with 64D descriptors and modified SURF with 16D descriptors. The vision system is mounted on a tripod and moved in a counterclockwise direction. The tripod passes through all nine ground truth positions in each circulation. The SLAM system with conventional SURF successfully ran for only 4307 frames. Figures 14 and 15 compare the results of the SURF method and modified SURF method, respectively. Figure 16 presents the map size at each frame. The map size of the modified SURF method increases to about 220 features at the first loop-closure and then stabilizes from the second to the fourth loop-closure. On the other hand, the map size of the conventional SURF method increases continuously after the first loop-closure. For the modified SURF method, the traveling time for the first to fourth circulation around the circular path was 98.6 s, 55.5 s, 51.8 s, and 50.3 s, respectively. The time to reach first loop-closure was longer than that to reach second loop-closure because the vision system moved slowly in order to provide enough time for SURF computation. In first circulation, the SLAM system had to spend time on computing the corresponding SURFs in right camera and also on initializing all new features. Table II lists the state estimation of ground truth points for the case of modified SURF with 16D descriptors. The first three columns of the table are ground truth locations of ground truth points, and the last three columns are the estimated locations of those points. The ± variation values indicate the standard deviations in the estimated values. The

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

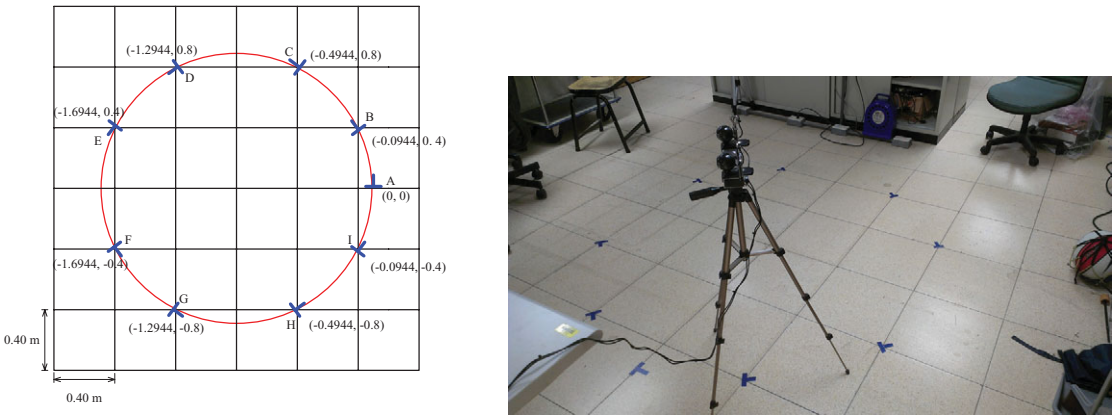


Fig. 13. (Colour online) Ground truth locations.

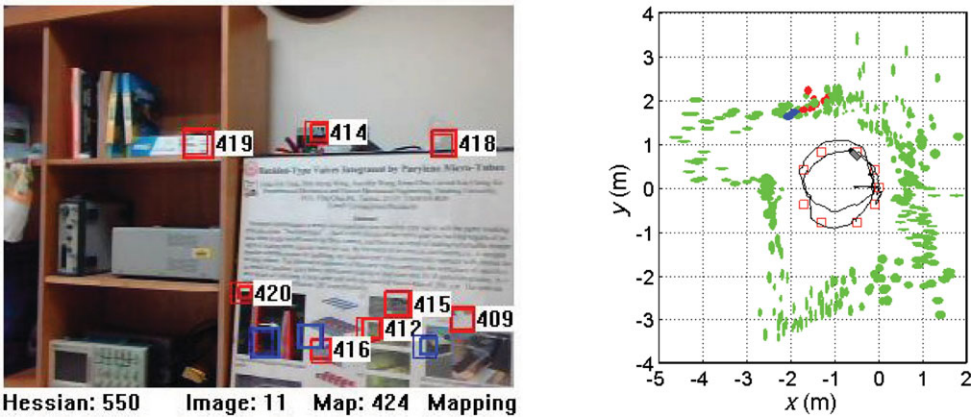


Fig. 14. (Colour online) Conventional SURF with 64D descriptors.

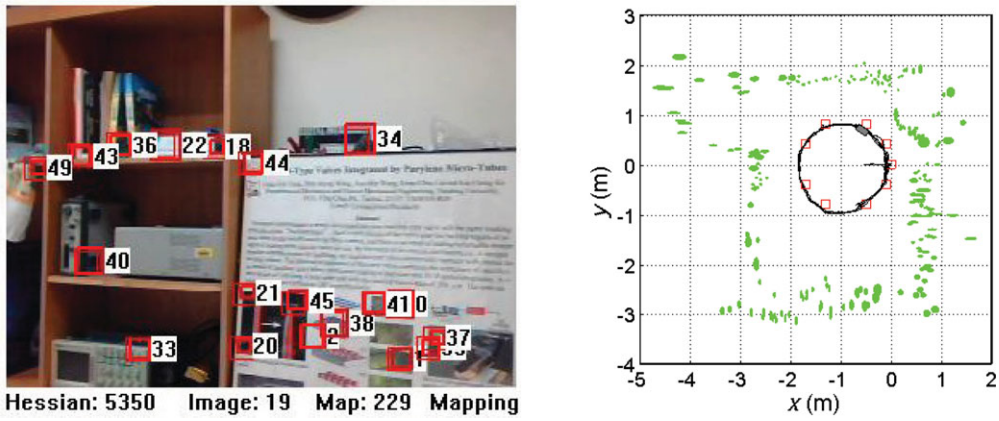


Fig. 15. (Colour online) Modified SURF with 16D descriptors.

experimental results show that the modified SURF with 16D descriptors provides a more stable and accurate state estimation compared to conventional SURF with 64D descriptors.

5.3. Robot SLAM

In this experiment, the camera is hand carried counterclockwise around the bookshelves (7 m × 0.5 m floor dimension) in the library. Figure 17 plots the resultant map and the camera pose estimation. This figure presents the estimated states of the camera and landmarks in a 3D map plot. The ellipses

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

Table II. State estimation of ground truth points.

	Ground truth (m)			Estimated (m)		
Points	x	y	z	$x(\mu \pm \sigma)$	$y(\mu \pm \sigma)$	$z(\mu \pm \sigma)$
A	0.00	0.0	0.0	-0.0962 ± 0.0190	-0.0474 ± 0.0265	0.1043 ± 0.1746
B	-0.0944	0.4	0.0	-0.1505 ± 0.0197	0.3397 ± 0.0276	0.0427 ± 0.0797
C	-0.4944	0.8	0.0	-0.5553 ± 0.0099	0.7025 ± 0.0116	0.0245 ± 0.0358
D	-1.2944	0.8	0.0	-1.3569 ± 0.0053	0.7156 ± 0.0355	-0.0041 ± 0.0279
E	-1.6944	0.4	0.0	-1.7541 ± 0.0078	0.3326 ± 0.0158	-0.0097 ± 0.0170
F	-1.6944	-0.4	0.0	-1.7362 ± 0.0127	-0.4629 ± 0.007	0.0003 ± 0.0348
G	-1.2944	-0.8	0.0	-1.3709 ± 0.0036	-0.8906 ± 0.0119	0.0185 ± 0.0385
H	-0.4944	-0.8	0.0	-0.6263 ± 0.0189	-0.8345 ± 0.0406	0.0772 ± 0.0606
I	-0.0944	-0.4	0.0	-0.1903 ± 0.0077	-0.4593 ± 0.0170	0.1355 ± 0.0775

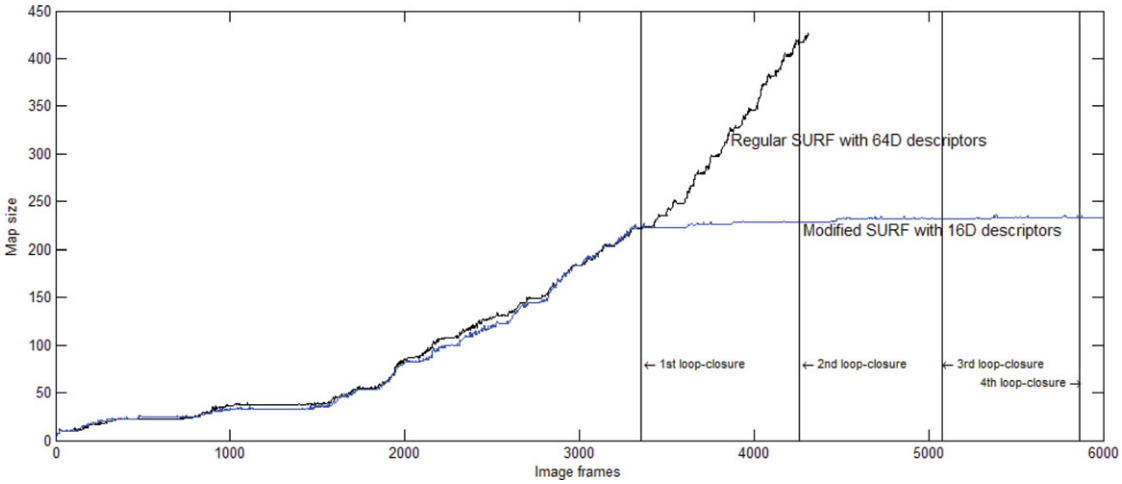


Fig. 16. (Colour online) Map size.

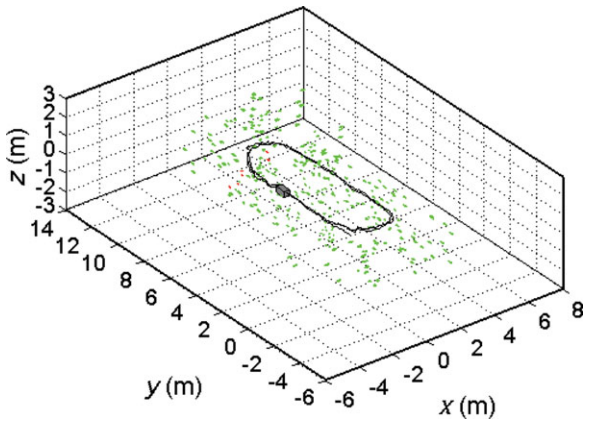


Fig. 17. (Colour online) Navigation performance of robot SLAM around library bookshelves.

in the figure indicate the uncertainty of the landmarks obtained from the extracted image features. The solid line represents the trajectory of the free-moving binocular vision. The rectangular box represents the left camera of the system, which is the location of camera frame $\{C\}$. From the first image frame, the camera circles the bookshelves two times. The SLAM system also starts from the first image frame and captures image features with unknown positions. After these features are initialized, they are stored as landmarks in the map. As the camera circles the bookshelves, the SLAM system concurrently builds the environment map and estimates the camera pose.

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press



Fig. 18. (Colour online) The seventh image frame: when $D_{\text{threshold}} = 5000$, 10 features are detected and stored in the map.

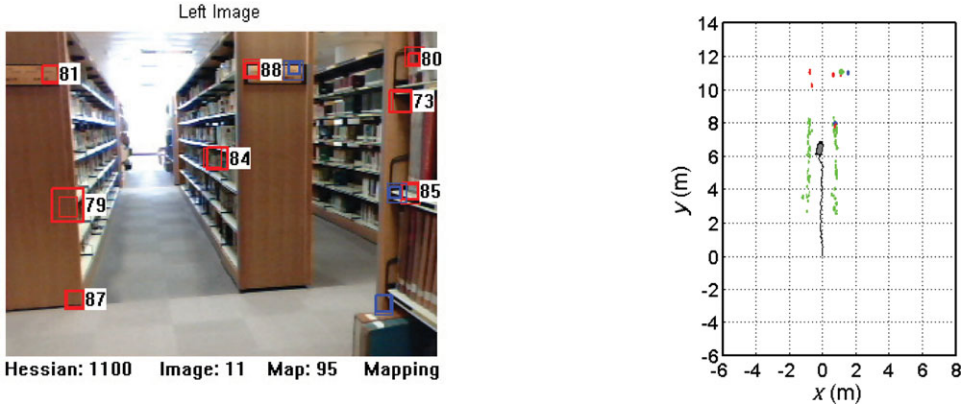


Fig. 19. (Colour online) Frame 645: when $D_{\text{threshold}} = 1100$, 11 features are detected. The map has accumulated 95 landmarks.

Figures 18–25 show the image frames obtained in the experiments in further detail. For each figure, the captured image is shown in the left panel and the top-view plot is depicted in the right panel. The (blue) square marks in the left panel of the figures indicate the landmarks with unstable matching rates after extraction from the captured images while the (red) square marks represent the landmarks with a stable matching rates and image depths. The right panel of the figures shows a 2D plot of the estimated states of the camera and landmarks. The red (dark) ellipses represent the uncertainty of the landmarks that have known image depths, and the green (light) ellipses denote the uncertainty of the landmarks that have unknown image depths. The SLAM task begins from the first frame. Figure 18 presents the SLAM task at the seventh frame. The left panel of the figure shows the 10 features detected in the captured image and stored in the map. The threshold value of the determinant of the Hessian matrix is 5000 at this frame. Figures 18–23 depict the first circulation of the camera around the bookshelves. During the circulation, the system continuously accumulates static features for use as landmarks in the map. The 664th frame in Fig. 20 shows how the map management system detected and discriminated a moving object (a student walking in a library) from stationary objects. The 1750th frame in Fig. 23 is the end of the first camera circulation. At this location, the camera comes to the place it had visited before and the trajectory loop is closed. Some of the old landmarks, e.g., those below 100, are detected again, and the covariance of the state vector is gradually reduced. Frames 2050 and 2560 in Figs. 24 and 25, respectively, describe the second circulation around the bookshelves. The landmark numbers under 100 are revisited and successfully detected in the 2560th frame by the SLAM system.

Figure 26 depicts the map size and the sampling frequency (Hz) at each frame. The map size increases to about 320 features at the first loop-closure and then stabilizes. That is, the SLAM relies on the map with old landmarks for localization when it revisits the same environment. The highest

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

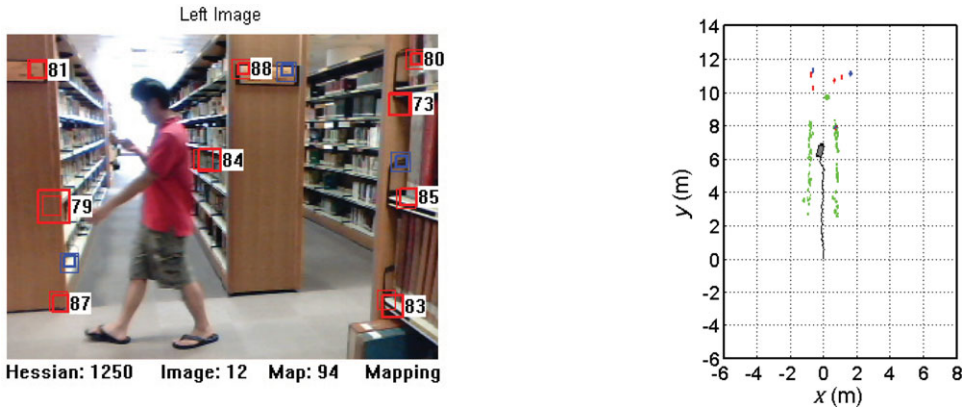


Fig. 20. (Colour online) Frame 664: $D_{\text{threshold}} = 1250$, and 12 features are detected. The map has 94 landmarks.

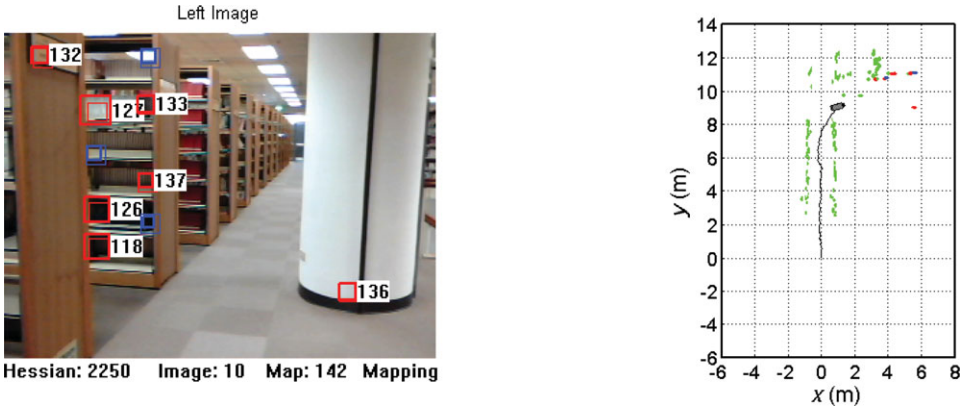


Fig. 21. (Colour online) Frame 920: the camera begins to circulate around the bookshelves. In this frame, $D_{\text{threshold}} = 2250$, and 10 features are detected. The map has 142 landmarks.

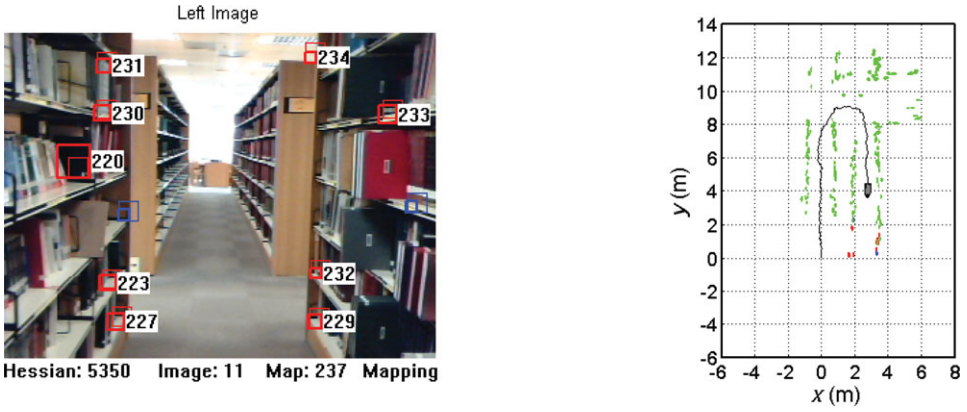


Fig. 22. (Colour online) Frame 1364: since the number of available features is increased, $D_{\text{threshold}}$ increases to 5350. Eleven features are detected. The map size increases to 237.

sampling frequency in the figure is about 20 Hz when the map size is small. As the map size increases, the sampling frequency remains as high as 7 Hz. The traveling time for the first and second circulation around the bookshelves was 217 s and 157 s, respectively. Figure 27 plots the deviations of the camera pose in the *xyz*-axis. The figure shows that, during the SLAM task, the average pose deviation is less than 2 cm, and the highest peak is no more than 4 cm.

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

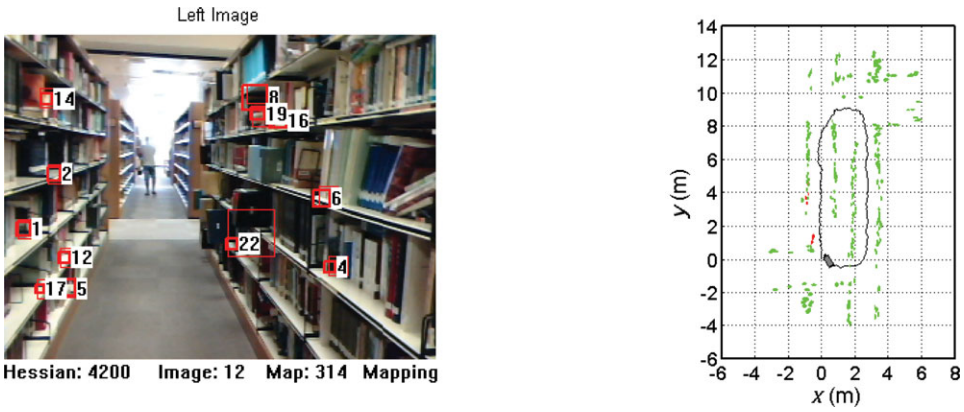


Fig. 23. (Colour online) Frame 1750: the camera returns to a previous scene. The feature numbers below 100 are detected again.

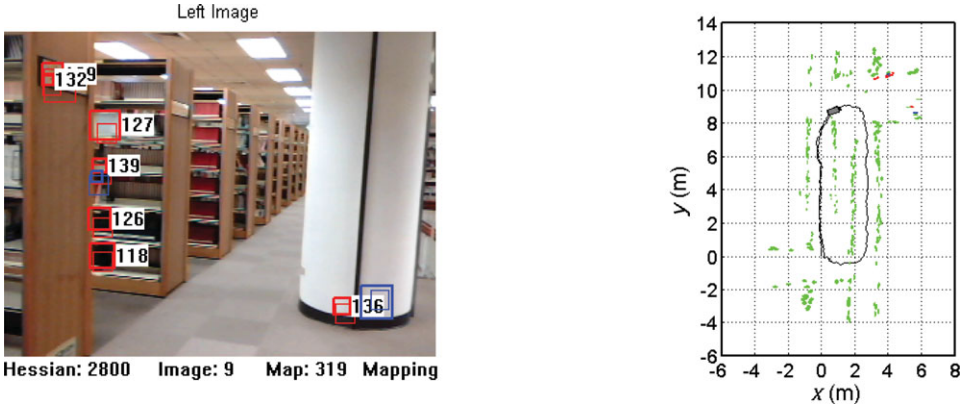


Fig. 24. (Colour online) Image in frame 2050: $D_{\text{threshold}} = 2800$, and nine features are detected. The map has 319 landmarks.

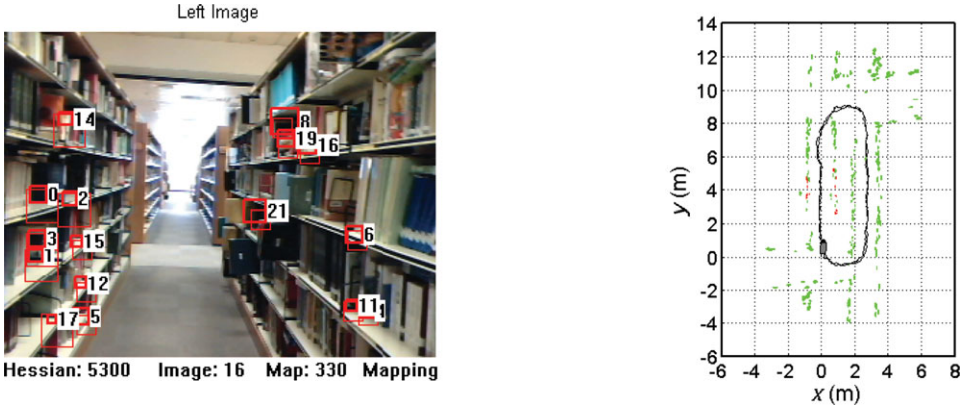


Fig. 25. (Colour online) Image in frame 2560: at the end of the second circulation of the camera, the feature numbers below 100 are detected again.

6. Conclusion

This study developed a robot mapping procedure for using SURF algorithms for robot SLAM. Based on the invariant properties of SURF, landmark scales and orientations are automatically recovered. For application in dynamic environments, this study proposed many modifications of the SURF algorithms including feature orientation representation, vector dimension of feature description, and number of detected features in an image. Experiments performed in an actual system with binocular vision sensors validated the feasibility and effectiveness of proposed algorithms. In performance

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

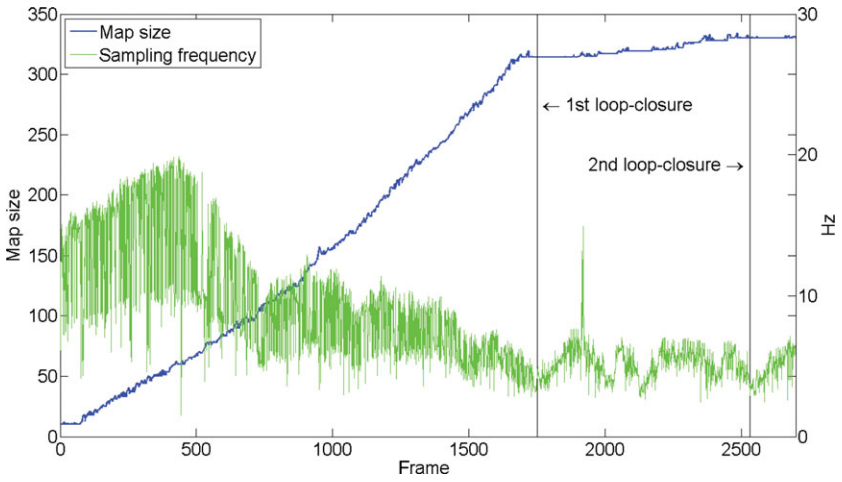


Fig. 26. (Colour online) Map size and sampling frequency *versus* the number of image frames.

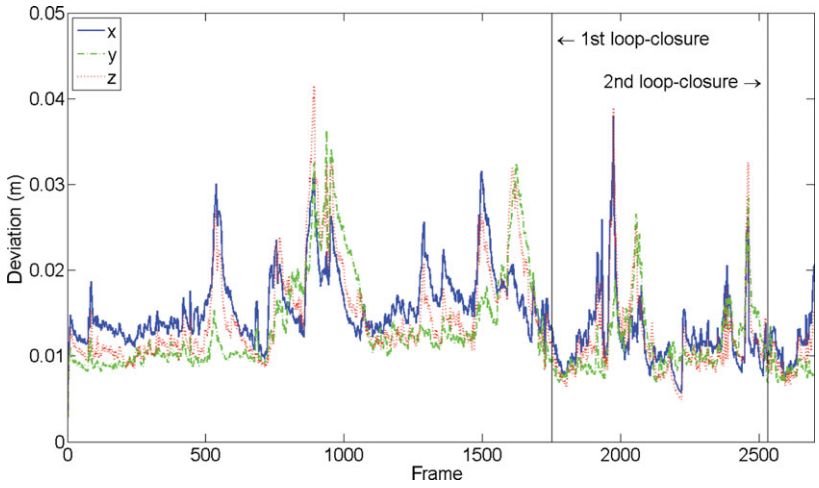


Fig. 27. (Colour online) Deviation in camera position *versus* image frame.

comparisons with conventional SURF algorithms, the first experiment showed that the modified SURF method improves feature repeatability by 10%. Ground truth experiments also showed that the modified SURF robustly represents visual landmarks during SLAM tasks. In contrast, conventional SURF did not provide a stable representation of the landmarks. Additionally, a large-area SLAM was successfully implemented using EKF SLAM with the modified SURF algorithms.

Acknowledgements

The authors thank the National Science Council of Taiwan for partially supporting this research under grant no. NSC101-2221-E-032-017 and NSC101-2632-E-032-001-MY3. Ted Knoy is appreciated for his editorial assistance.

References

1. J. Artieda, J. M. Sebastian, P. Campoy, J. F. Correa, I. F. Mondragón, C. Martínez and M. Olivares, "Visual 3-d slam from uavs," *J. Intell. Robot. Syst.* **55**, 299–321 (2009).
2. H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, "SURF: Speeded-up robust features," *Comput. Vis. Image Underst.* **110**, 346–359 (2008).
3. C. Cocaud and T. Kubota, "SURF-Based SLAM Scheme Using Octree Occupancy Grid for Autonomous Landing on Asteroids," *Proceedings of the International Symposium on Artificial Intelligence, Robotics and Automation in Space,* Sapporo, Japan, (2010) pp. 275–282.

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press

4. A. J. Davison, I. D. Reid, N. D. Molton and O. Stasse, "MonoSLAM: Real time single camera SLAM," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 1052–1067 (2007).
5. A. Gil, O. M. Mozos, M. Ballesta and O. Reinoso, "A comparative evaluation of interest point detectors and local descriptors for visual SLAM," *Mach. Vis. Appl.* **21**, 905–920 (2010).
6. C. Harris and M. Stephens, "A Combined Corner and Edge Detector," *Proceedings of the 4th Alvey Vision Conference*, University of Manchester, Manchester, UK (1988) pp. 147–151.
7. N. Karlsson, E. di Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian and M. E. Munich, "The vSLAM Algorithm for Robust Localization and Mapping," *Proceedings of the IEEE International Conference on Robotics and Automation*, Barcelona, Spain (2005) pp. 24–29.
8. T. Lindeberg, "Feature detection with automatic scale selection," *Int. J. Comput. Vis.* **30**(2), 79–116 (1998).
9. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.* **60**(2), 91–110 (2004).
10. A. C. Murillo, J. J. Guerrero and C. Sagues, "SURF Features For Efficient Robot Localization with Omnidirectional Images," *Proceedings of the IEEE International Conference on Robotics and Automation*, Roma, Italy (2007) pp. 3901–3907.
11. L. M. Paz, P. Pinies, J. D. Tardos and J. Neira, "Large-scale 6-DOF SLAM with stereo-in-hand," *IEEE Trans. Robot.* **24**(5), 946–957 (2008).
12. G. Shakhnarovich, T. Darrell and P. Indyk, *Nearest-Neighbor Methods in Learning and Vision* (The MIT Press, MA, 2005).
13. P. A. Viola and M. J. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii (2001) pp. 511–518.
14. C. Y. Wu and L. C. Fu, "An Integrated Robotic vSLAM System to Realize Exploration in Large Indoor Environment," *Proceedings of the CACS International Automatic Control Conference*, Taichung, Taiwan (2007).

https://doi.org/10.1017/S0263574713000830 Published online by Cambridge University Press