2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Mobile Robot Localization Using Fusion of Object Recognition and Range Information

Byung-Doo Yim, Yong-Ju Lee, Jae-Bok Song and Woojin Chung

Abstract— **Most present localization algorithms are either range or vision-based. In many environments, only one type of sensor cannot often ensure successful localization; furthermore, using low-priced range sensors instead of expensive, but accurate, laser scanners often lead to poor performance. This paper proposes an MCL-based localization method that robustly estimates the robot pose with fusion of the range information from a low-cost IR scanner and the SIFT based visual information gathered using a mono camera. With sensor fusion, the rough pose estimation from range-based sensors is compensated by the vision-based sensors and slow object recognition can be overcome by the frequent update of the range information. In order to synchronize the two sensors with different bandwidths, the encoder information gathered during object recognition is exploited. This paper also suggests a method for evaluating localization performance that is based on the normalized probability of a vision sensor model. Various experiments show that the proposed algorithm can estimate the robot pose reasonably well and can accurately evaluate the localization performance.**

I. Introduction

Localization is a method for estimating the pose of a robot with an environmental map and information from sensors mounted on the robot. Localization is a fundamental and important task for the autonomous mobile robot. Range sensors, such as laser and IR scanners have been extensively used for global localization. However, when only range sensors are employed, the estimation error of the robot's pose increases when in dynamic or cluttered environments. It is also difficult to find an accurate robot pose when the robot is placed in a simple environment like a hallway. On the other hand, a vision sensor usually provides more information than a range sensor, and it provides good performance at a low cost. Therefore, localization using vision sensors has drawn much attention in recent years. However, most algorithms on vision-based localization suffer from shortcomings in that their implementation takes longer than range-based localization because of the computation time needed for extracting feature information from the camera image.

In the past decade, substantial efforts have been directed toward the development of vision-based global localization. In topological Markov localization [1], the input image was compared to the images stored at each node of a topological map and then the node at which the robot was located was found by Markov localization. In this method, however, it is difficult to get an accurate robot pose if the robot is between nodes. Markov localization uses the ceiling information [2] [3], which led to successful localization even when people surrounded the robot. However, in an environment with a low ceiling, the robot could not collect sufficient information for localization from the camera image because it covered only a small portion of the ceiling. A Vision-based SLAM was proposed that used the Scale Invariant feature transform (SIFT) algorithm [4] based on a stereo vision [5] [6]. This approach required a large amount of memory because it must store all keypoints of the entire environment and it uses three cameras (a mono camera and a stereo camera).

Localization using relatively cheap sensors is important from a practical point of view, but localization with low-priced sensors seldom provides good localization performance in various environments due to inaccurate sensor measurements. Either the range-based or vision-based scheme alone cannot overcome these sensor limitations; therefore, sensor fusion based localization should be implemented to compensate for shortcomings of each sensor. This paper proposes the global localization algorithm based on the the fusion of the range information from a low-cost IR scanner and the visual information from a mono camera. The proposed localization scheme is mainly based on the Monte Carlo Localization (MCL) algorithm [7]. Dependable navigation is possible since the relatively poor range accuracy from an IR scanner can be compensated through vision-based localization and slow object recognition can be overcome by the frequent update of the range information.

One problem involved in the fusion of range and visual data is their different processing times. That is, the range information has a higher update rate than the visual information because object recognition based on SIFT feature extraction requires a long computation time, especially when the object has many features. In this paper, the data from the two sensors are synchronized by compensating for the time delay caused by the slow vision-based localization by using the encoder information.

This research was conducted byr the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

Byung-Doo Yim is with the Dept. of Mechatronics, Korea University, Seoul, Korea (e-mail: ybd1021@korea.ac.kr).

Yong-Ju Lee is with the Dept. of Mechanical Eng., Korea University, Seoul, Korea (e-mail: yongju_lee@korea.ac.kr).

Jae-Bok Song is a Professor of the Dept. of Mechanical Eng., Korea University, Seoul, Korea (Tel.: +82 2 3290 3363; fax: +82 2 3290 3757; e-mail: jbsong@korea.ac.kr)

Woojin Chung is an Assistant Professor of the of the Dept. of Mechanical Eng., Korea University, Seoul, Korea (Tel.: +82 2 3290 3375; e-mail: smartrobot@korea.ac.kr)

Another issue of global localization is the evaluation of localization performance. The capability of detecting and recovering from localization failure is essential for autonomous navigation, because no localization algorithm guarantees its success at all times. The MCL algorithm uses random samples to cope with the kidnapped robot problem and localization failure [8], but if the number of random samples is not sufficient, the time required to detect localization failure may be quite long. This paper proposes a scheme that evaluates the localization performance based on the normalized probability of the vision sensor model. If the localization performance is regarded as localization failure, then the proposed scheme can recover the robot pose using the recognized object.

The remainder of this paper is organized as follows: Section II presents the vision sensor model and the range sensor model. Section III presents the fusion of the two sensor models for MCL. Section IV shows the detection and recovery from localization failure. Finally, conclusions are outlined in section V.

II. Sensor Models

In this research, the range and vision sensors are fused together for improved localization of a mobile robot. Instead of a laser scanner, which is very accurate but expensive, an IR scanner is used as the main range sensor. The IR scanner generates a vector of 121 range values with a resolution of 1.8°. An inexpensive mono camera is also employed as the main vision sensor instead of a stereo camera, which is expensive but can provide range information. Objects are recognized by the well-known SIFT algorithm to extract the visual features. The sensor model for each sensor is required for probability update of random samples (i.e., candidates for the robot pose) used in MCL.

A. Range sensor model

In the range sensor model, the probability of samples is updated according to the difference between the range data measured by the IR scanner and those computed from the sample pose on the map, as shown in Fig. 1. That is, if the robot pose at time t is denoted as x_t, the probability of sample i ($i = 1, \ldots, N$) is updated by

$$p_{ir}^{(i)}(z_t \mid x_t) = \frac{1}{\sum_{k=1}^{k_t} (z_t(k) - d_t^{(i)}(k))^2} \tag{1}$$

where $z_t(k)$ represents the k-th value of the range data measured at time t ($k = 1, .., 121$), and $d_t^{(i)}(k)$ is the k-th value of the range data computed from sample i on the map. Although the IR scanner provides a total of 121 range data, only the data less than 4m (k_t is the total number of such data) are used in Eq. (1) because the data exceeding 4m are found to be incorrect.

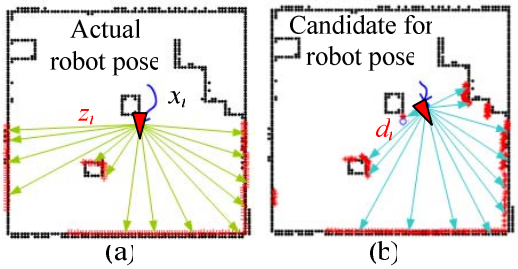

Fig. 1 Range data; (a) measured by the actual sensor, and (b) computed from candidate robot pose.

B. Vision sensor model

In the vision sensor model, the probability is updated according to the difference between the measured range and angle to the recognized object and those computed from the samples on the map. The center of the object is selected as a point representing this object because the object has its own image size on the image plane. The relative range and angle from the robot to the object are extracted with respect to this center point. Therefore, robust and accurate extraction of the center point of the recognized object is very important in minimizing the localization error. The following affine transform, which calculates the geometrical relations between the recognized object and the object stored in the database, is used to extract the accurate center point [4].

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{2}$$

where the vector $[t_x, t_y]^T$ is associated with the translation and the parameter m_i ($i = 1, .., 4$) with the 3D rotation. The vector $[x_i, y_i]^T$ is the keypoint relative to the image stored in the database, and $[u_i, v_i]^T$ is the keypoint extracted in the current image, as shown in Fig. 2.

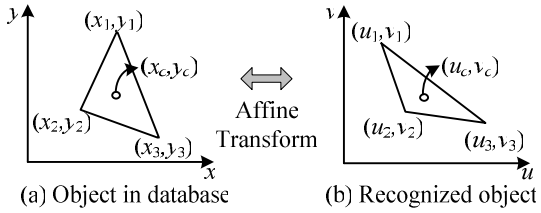

Fig. 2 Example of affine transform.

Equation (2) is rearranged to compute the 6 parameters as follows:

$$\begin{bmatrix} x_1 & y_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_1 & y_1 & 0 & 1 \\ x_2 & y_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_2 & y_2 & 0 & 1 \\ & & \cdots\cdots & & & \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ \vdots \end{bmatrix} \tag{3}$$

The parameters can be computed by inversion of the 6 by 6 matrix if only 3 pairs of matched keypoints are given, as shown in Fig.2. For more than 3 pairs, the pseudo inverse matrix is used to yield the 6 parameters. Once the parameters are identified, the center point of the recognized object, (u_c, v_c), can be computed by Eq. (2) from that of the object in the database, (x_c, y_c).

To obtain the relative angle to the recognized object from the robot, we need to transform the extracted center point of the object relative to the image plane into that in the robot frame. This angle at time t, denoted as θ_t^{obj}, is given by

$$\theta_t^{obj} = \tan^{-1}\left(\frac{(w_{image}/2)-u_c}{(w_{image}/2)}\cdot\tan(\theta_{fov}/2)\right) \quad (4)$$

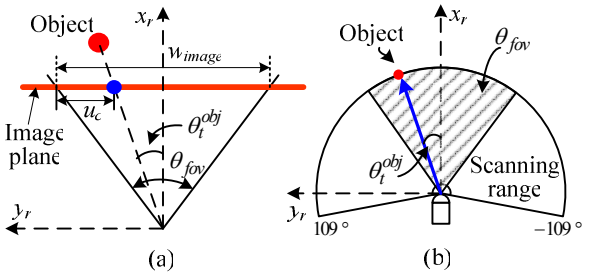

Fig. 3 Extraction of visual feature; (a) the center point of the object in the image, (b) detection range of IR scanner.

where (x_r, y_r) of Fig. 3(a) represents the robot frame; u_c is the coordinate of the center point relative to the image frame, w_{image} is the number of pixels (e.g., 320 pixels) of the image plane in the u axis of the image frame, and θ_{fov} is the camera's field of view.

In this research, the IR scanner is used to compensate for the mono camera's inability to provide the range information. As shown in Fig. 3(b), the camera's field of view is always included in the scanning range of the IR scanner. Therefore, the range reading corresponding to the angle θ_t^{obj} is used as the range to the object's center point. A more accurate range value can be obtained by interpolating between the two adjacent range data that are separated by an angle of 1.8°, which happens to be the resolution of an IR scanner.

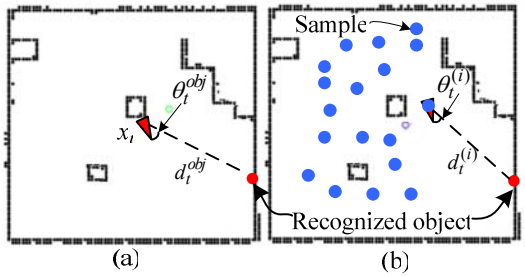

Fig. 4 Range and relative angle to the recognized object; (a) measured from the robot, and (b) computed from sample i.

In the vision sensor model shown in Fig. 4, the range (d_t^{obj}) and relative angle to the recognized object from the robot (θ_t^{obj}), are compared with $d_t^{(i)}$ and $\theta_t^{(i)}$, which are the range and relative angle to the object from the robot computed from sample i on the map, respectively. Based on the difference between the measured and the computed results, the probabilities are updated as follows:

$$p_d^{(i)}(z_t \mid x_t) = \eta_r \frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\left(-\frac{1}{2}\frac{(d_t^{obj}-d_t^{(i)})^2}{\sigma_d^2}\right) \quad (5)$$

$$p_\theta^{(i)}(z_t \mid x_t) = \eta_\theta \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left(-\frac{1}{2}\frac{(\theta_t^{obj}-\theta_t^{(i)})^2}{\sigma_\theta^2}\right) \quad (6)$$

where $p_d^{(i)}(z_t \mid x_t)$ and $p_\theta^{(i)}(z_t \mid x_t)$ are the probabilities associated with the respective range and relative angle, and η_r and η_θ are the normalizing constants for the range and angle, respectively. Each sensor model has a Gaussian distribution with the mean of d_t^{obj} and θ_t^{obj}, and a variance of σ_d^2 and σ_θ^2. For each sensor model, the overall vision sensor model is given by

$$p_v^{(i)}(z_t \mid x_t) = p_d^{(i)}(z_t \mid x_t) \times p_\theta^{(i)}(z_t \mid x_t) \quad (7)$$

III. Fusion of Range and Vision Sensors

Vision-based localization can generally give more effective localization performance than range-based localization provided there are many objects with visual features in the environment. However, because only a small number of objects can be used as visual features in normal indoor environments, vision-based localization alone is not sufficient to provide satisfactory localization performance in most environments. Thus, if the recognized objects cannot be found at the current robot pose, only the range sensor model can be used to update the probability of samples as follows:

$$p(z_t \mid x_t) = p_{ir}(z_t \mid x_t) \quad (8)$$

where $p_{ir}(z_t|x_t)$ is the range sensor model given by (1). If the vision sensor recognizes any object, the range sensor model and the vision sensor model are fused to update the probability of samples.

$$p(z_t \mid x_t) = p_v(z_t \mid x_t)\, p_{ir}(z_t \mid x_t) \quad (9)$$

where $p_v(z_t|x_t)$ is the vision sensor model given by (7).

It is important that the data from the IR scanner and the vision sensor are fused in a synchronous fashion. However, in contrast to the relatively fast response of an IR scanner, vision-based object recognition often requires a rather long processing time. Due to this delay, the information obtained upon completion of the object recognition actually reflects the environment information at the beginning of object recognition. For synchronization purposes, the range data measured at the start of object recognition must be fused with

the visual data at the end of object recognition, as shown in Fig. 5. As the processing time for object recognition increases, several sets of recent range data should be discarded for synchronization with the vision data, as shown in Fig. 5. Thus, the overall update rate of the sample probability becomes low, which leads to an increasing failure rate of localization due to lack of the most recent environment information.

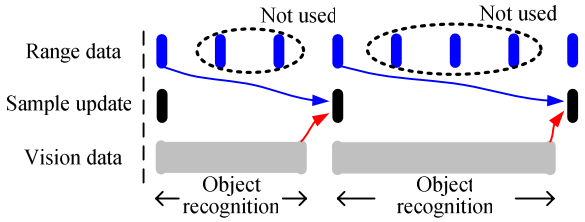

Fig. 5 Sensor fusion with loss of range information.

In order to cope with this problem, the range data and the vision data are used separately in this research. That is, the range data continue to be used to update the probability of the samples while object recognition is in process. Sensor fusion is conducted only when object recognition is completed and the visual data are available.

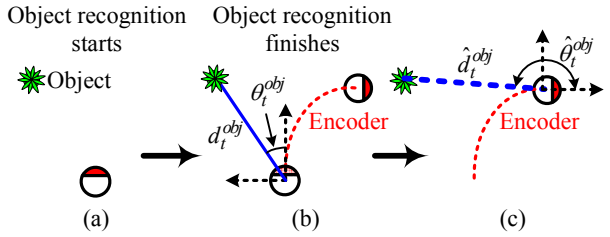

Fig. 6 Relation between robot pose and object when object recognition starts and finishes.

As shown in Fig. 6, the observation $(d_t^{obj}, \theta_t^{obj})$ obtained at the end of object recognition is actually based on the previous robot pose because the robot is moving while object recognition is in process. Therefore, this observation must be compensated to reflect the current robot pose. This compensation can be performed using the encoder data, by assuming that a change in the robot pose estimated by the encoder data over a short period of time is relatively accurate. This compensation based on the encoder data is applied to all samples, as shown in Fig. 7(d).

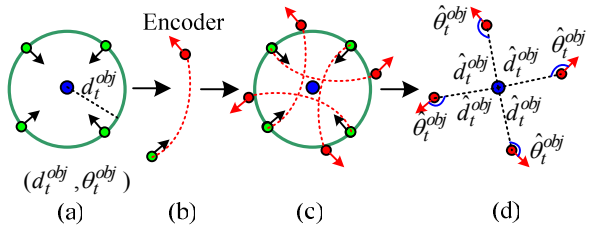

Fig. 7 Prediction of relation between robot pose and object using encoder data at the end of object recognition.

The uncertainty in the encoder data should be considered in computation of $(\hat{d}_t^{obj}, \hat{\theta}_t^{obj})$ as follows.

$$\begin{aligned} \sigma_d + \alpha_1|\Delta d| + \alpha_2|\Delta\theta| &= \sigma_{\hat{d}} \\ \sigma_\theta + \alpha_3|\Delta d| + \alpha_4|\Delta\theta| &= \sigma_{\hat{\theta}} \end{aligned} \tag{10}$$

where σ_d and σ_θ are the respective uncertainties in the range and the relative angle measured by the vision sensor, and Δd and $\Delta\theta$ are the translational and rotational motion of the robot during its object recognition, respectively. The parameters, α_1 and α_2 associated with $\sigma_{\hat{d}}$, and α_3 and α_4 with $\sigma_{\hat{\theta}}$, depend on the characteristics of the robot. These parameters depend on the robot. In the real experiments, σ_d and σ_θ were set to 0.2m and 3°, respectively. The parameters α_1, α_2, α_3 and α_4 were set to 0.1, 0.2m/deg, 0.5°/m and 2° respectively. It is noted that the uncertainty $\sigma_{\hat{d}}$ and $\sigma_{\hat{\theta}}$ increase with an increase in Δd and $\Delta\theta$.

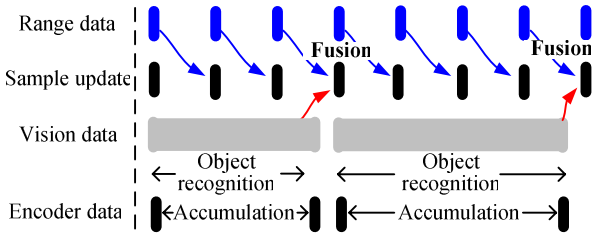

Fig. 8 Sensor fusion without loss of sensor information.

Various experiments were performed using a robot equipped with an IR scanner (Hokuyo PBS-03JN) and a mono camera (a normal web camera). As shown in Fig. 9(a), the experimental environment was 15m x 80m and consisted of a long hallway and several doors. The grid size of the grid map was 10cm. Figure 9(b) illustrates the objects used as visual landmarks for localization, and their positions are shown as red dots in Fig. 9(a).

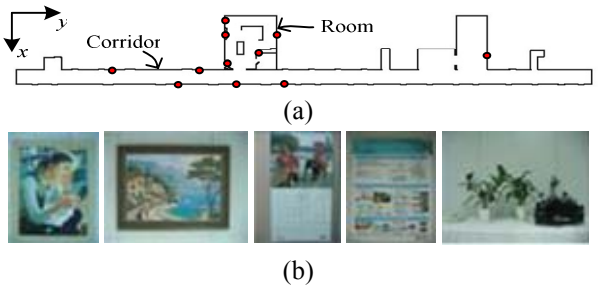

Fig. 9 (a) global map of experimental environment, and (b) objects used as visual landmarks.

11 visual landmarks were used in the experiments for global localization and were carried out in a room and a hallway. When the robot navigates through a hallway, the information from the range sensor is not sufficient for successful localization, which often leads to slow convergence of the samples. 5,000 samples were initially distributed throughout

the entire environment and these samples were converged to a small local area by continuously updating the probability of samples using the sensor information. Localization was determined to be completed once the sample variances become smaller than the pre-determined thresholds.

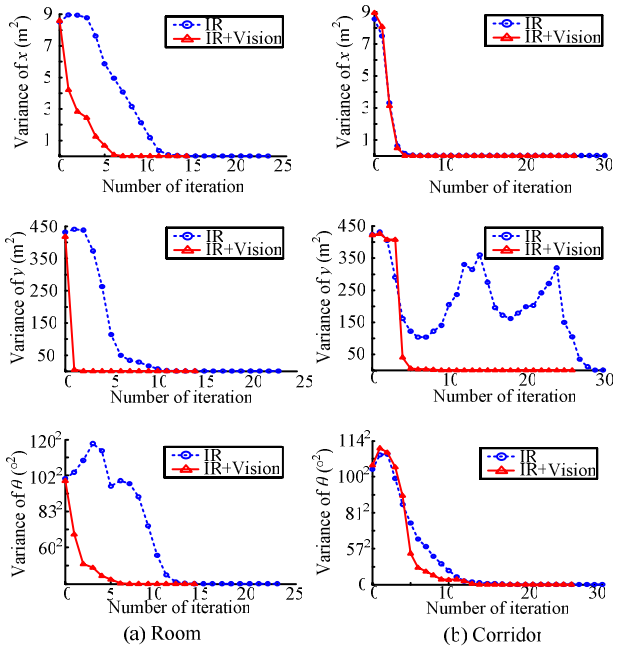

Fig. 10 Variance of sample position in x, y axis and θ in localization; (a) room, (b) hallway.

Suppose the robot was placed at a room at the beginning of MCL. As shown in Fig. 10(a), the sample variances converge to zero and thus the estimated robot pose can keep track of the actual pose reasonably well because enough environmental information can be obtained only from only the range sensor if a robot is located in a room. However, if the vision data are fused with the IR scanner data, the sample variances converge to zero more rapidly than with only the range data. In the fusion-based localization, fast convergence can be achieved once the objects are visually recognized. Note that the variance associated with the y-axis (i.e., along the hallway) is much larger than that with the x-axis because the range data in the y-axis is quite uncertain due to the limitations of the range sensor (4m in this experiment).

If a robot is located at a hallway at the beginning of MCL, the localization performance is generally worse than when a robot being is in a room. First of all, few geometric features can be collected by the range sensor because the geometric information in the hallway is quite similar from place to place. Furthermore, the IR scanner has a relatively short measuring range and the sample variances associated with the y-axis do not converge satisfactorily, sometime they diverge as shown in Fig, 11(b). Even in this case, however, if sensor fusion is conducted for localization, a better convergence of the sample variances can be achieved, thus resulting in successful localization.

In the proposed method of sensor fusion, the probability update of samples becomes more efficient in that all the data from the range and vision sensor can be used without any loss of data. In addition, the data from the vision sensor is the range and relative angle to the object from the robot. Therefore, if only the vision sensor is used to find the robot pose, at least two objects must be observed. With fusion of the vision sensor and range sensor, however, it is possible to find the robot pose even when only one object is recognized.

IV. Recovery from Localization Failure

If the gap between the given environment map and the sensor information becomes large in the localization of a mobile robot, the random samples in MCL can converge to give incorrect poses or divergence, thus resulting in localization failure. Other sources of localization failures are caused by inaccurate encoder data due to slippage during navigation and the kidnapped robot problem. In case of localization failure, it is important to detect this failure and recover from it for dependable navigation.

To evaluate localization performance, the probabilistic vision sensor model, explained in section II, is adopted in this research. The probability associated with the vision sensor model depends on the uncertainty $\sigma_{\hat{d}}$ and $\sigma_{\hat{\theta}}$ given by Eq. (10). It is difficult to select a fixed probability threshold to determine whether localization is successful because these uncertainties change at the end of each iteration of MCL. Therefore, a new criterion to evaluate the localization performance is proposed in this research.

Figures 11(a) and (c) depict the probability distribution associated with the range to the visual object. Both distributions are Gaussian with a mean of 2m, but their standard deviations (representing the uncertainty) are set to 0.2m for Fig. 11(a) and 0.4m for Fig. 11(c). Figure 9(e) and (g) depict the probability distribution associated with the relative angle. Both are Gaussian with a mean of 0°, but their standard deviations are 3° for Fig. 11(e) and 5° for Fig. 11(g). Figures 11(b), (d), (f) and (h) represent the probability distributions normalized by their respective maximum probabilities (indicated as a red dot in Fig 11(a), (c), (e) and (g)).

The evaluation of localization performance is illustrated in Fig. 11(b), (d), (f) and (h). Localization performance or quality can be classified into three cases. Case A, corresponding to the upper 30% of the normalized probability distribution, is regarded as successful localization. Case C, the lower 30%, is recognized as localization failure and the remainder, case B, is classified as a warning. For instance, suppose the computed range and angle are given by 2.3m and 10° from the sensor model given in Fig. 11(c) and 11(g), which corresponds to the normalized probability of 0.75 and 0.13, respectively. In this situation, localization is judged as a failure, because the normalized probability associated with the angle falls into region C even though the probability associated with the range is in region A.

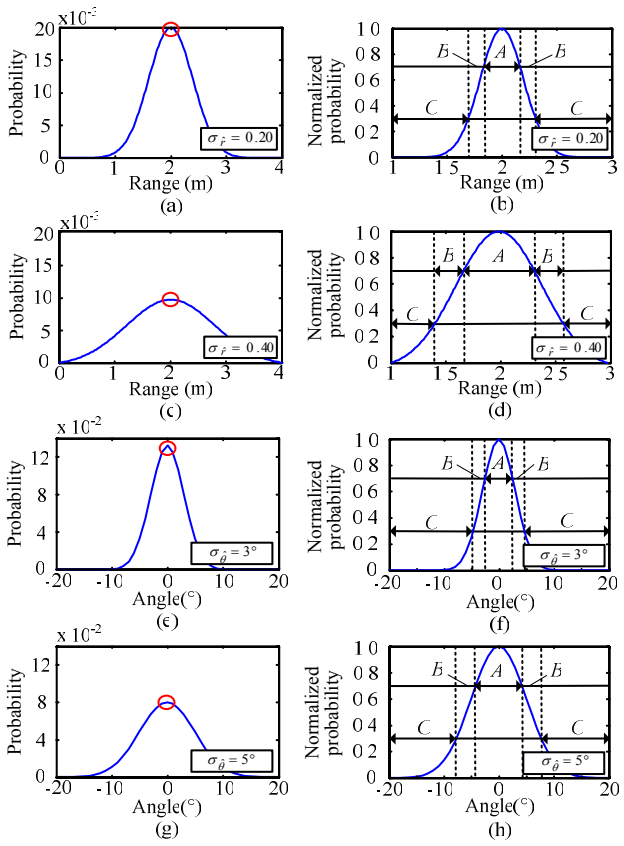

Fig. 11 Decision of the localization performance based on vision with uncertainty: (a), (c), (e) and (g) are the probability distributions of a visual feature, (b), (d), (f) and (h) are normalized probabilities and the localization performance, *A*: successful localization, *B*: warning, *C*: localization failure.

Judgment of localization failure requires that the normalized probability falls in region *C* 3 times in a row in order to avoid misjudgment due to false-matching in the object recognition based on SIFT features. Also, if the normalized probability falls into region *B* 10 times in a row, then localization is considered to fail.

The proposed algorithm for vision-based recovery from localization performance was investigated through various experiments. Figure 12 shows recovery from localization failure using the normalized probability. Once the robot is aware of a localization failure, it wanders to collect visual data while the range-based localization is in process. If an object is recognized by the vision sensor, it is needless to distribute the random samples for MCL on the entire empty area of the environment because the position of the recognized object can be approximated from the map information. As shown in Fig. 12(b), samples are mainly drawn near the circle with a radius of the measured range and centered at the recognized object. Obviously, this sample distribution is more efficient in localization than the uniform distribution covering the entire environment. Although visual features are rare in some environments, vision-based recovery of the robot pose is efficient and robust provided that visual features are available.

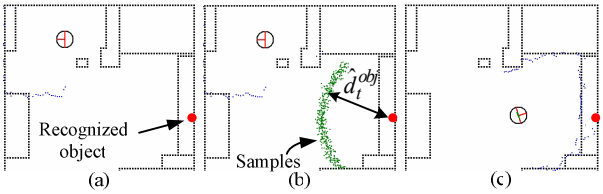

Fig. 12 Recovery from localization failure; (a) detection of localization failure, (b) distribution of samples near recognized visual feature, and (c) recovery of robot pose.

V. Conclusions

This paper proposes an efficient sensor fusion based localization algorithm, in which an IR scanner and a cheap web camera are used. From this research, the following conclusions have been drawn.

1) Sensor fusion based localization proposed here enables the samples in MCL to converge to the actual robot pose faster than either range-based or vision-based localization alone.
2) Although the processing time for object recognition takes a long time and is not periodic, the probability of samples can be updated at the speed of a range sensor using the proposed method.
3) The proposed algorithm for the evaluation of localization performance based on the vision sensor model works well to detect localization failure and recover from it.

Currently, research on the improvement in the accuracy and speed of the object recognition algorithm is under way.

References

[1] J. Kosecka and F. Li, "Vision based topological Markov localization," *Proc. of IEEE Int'l Conf. on Robotics and Automation*, vol. 2, pp. 1481-1486, 2004.
[2] S. Thrun et al., "Minerva: a second-generation museum tour-guide robot," *Proc. of IEEE Int'l Conf. on Robotics and Automation,* vol. 2, pp. 1999-2005, May, 1999.
[3] W.-Y. Jeong, K.-M. Lee, "CV-SLAM: A new Ceiling Vision-based SLAM technique," *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 3195-3200, Aug. 2005.
[4] D.G. Lowe, "Distinctive image features from scale invariant keypoints," *Int'l Journal of Computer Vision*, vol. 60, no 2, pp. 91-110, 2004.
[5] S. Se, D. Lowe and J. Little, "Mobile robot localization and mapping with uncertainty using scale invariant visual landmarks," *Int'l Journal of Robotics Research*, vol. 21, no.8, pp. 735–758, Aug. 2002.
[6] D.G. Lowe and S. Se, "Vision-Based global localization and mapping for mobile robots," *Proc. of IEEE Transactions on Robotics,* vol. 21, pp. 217-226, June, 2005.
[7] T.-B. Kwon, J.-H. Yang, J.-B. Song, W. Chung, "Efficiency Improvement in Monte Carlo Localization through Topological Information," *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Oct. 2006.
[8] S. Thrun, W. Burgard and D. Fox, "Probability Robotics," *MIT press*. 2005, ch. 8.