

# Map Building, Localization, and Object Detection for Indoor Mobile Robot Navigation Using Both Range and Vision Sensors

Tae-Bum Kwon, Jae-Bok Song, Munsang Kim

**Abstract**— In order to navigate in the real environment, lots of navigation techniques are required, such as mapping, localization, path planning, obstacle avoidance, and so on. The performance of these techniques can be influenced by the sensing ability and the algorithm. In this research, both range and vision sensors were used together to improve the sensing ability limited by sensor types, and three essential and important tasks were studied with these sensors. First, SLAM (Simultaneous Localization And Mapping) was investigated to build a map for unknown environments. Two types of features, the lines extracted from the range data and the objects recognized using the visual information, were exploited to improve the robustness of SLAM and the utility of the map. This SLAM could generate the grid map on which the positions of the recognized objects were registered. Second, localization, one of the most important tasks for a mobile robot, was studied. The map built by SLAM was used and the important problem was how to estimate the robot pose with fusion of the range information and the object information which were frequently and slowly updated respectively. In order to synchronize the two sensors with different bandwidths, the encoder information gathered during object recognition was exploited, and therefore the proposed localization algorithm could estimate the robot pose reasonably well. Finally, the object detection method was developed to avoid and interact with some objects, such as a table and a chair, which were very difficult to recognize using only a range sensor. In the proposed scheme, a robot detects the candidates for legs using the range data, generates many possible candidates for object pose (position and orientation) using the model information, and selects the best candidate by evaluating all candidates using visual information. The detected objects could then be mapped and exploited during navigation. Various experiments in the real environment showed the proposed schemes were effective in map building, localization and object detection with both range and vision sensors.

**Index Terms**— Mapping, Localization, SLAM, Object Detection.

## I. INTRODUCTION

In order to navigate in the real environment, several navigation techniques are required, such as mapping, localization, path planning, obstacle avoidance, and so on. Almost all techniques are based on the robot's perception. For example, a robot maps the perceived environment, estimates its pose using the sensed environment, and plans the path which

detours around the recognized obstacle. Therefore, perception is the most basic and important ability.

Three factors can influence the performance of perception; the sensing ability, the type of obstacles or environment, and the perception scheme. The sensing ability is limited by the sensor type. For example, a laser scanner offers very accurate and long-range data, but a sonar sensor gives less accurate and shorter range data than a laser scanner. The type of obstacles and the sensing ability are also closely related. For example, a range scanner can sense 2D obstacles, whereas a stereo vision sensor can sense 3D obstacles. These two factors, sensing ability and obstacle types, are related to the hardware and environment and not easy to deal with. Therefore, the perception scheme, or algorithm is important and should be improved to overcome the limitations of the sensing ability and type of obstacles.

Mapping, localization, and object detection are the most important and essential abilities for mobile robot navigation. These algorithms use the perception results, and in this research, both range and vision sensors are exploited to improve their performance. The first issue is the SLAM (Simultaneous Localization And Mapping) that is the process of building a map of the unknown environment and localizing a robot relative to this map at the same time. The majority of the existing SLAM methods for a mobile robot use a range sensor as main sensor [1][2][3]. A 2D range sensor, however, has a drawback that it can extract the limited types of features. A vision sensor also has a drawback that feature extraction time is relatively long while it can obtain much more information than range sensors. V-SLAM (Vision-based Simultaneous Localization and Mapping) and SIFT (Scale Invariant Feature Transform) are the most excellent examples [4][5].

This paper proposes an improved SLAM method which uses both the range sensor-based feature and the vision sensor-based feature. The most popular feature which can be extracted by a range sensor is the line feature, but it cannot be reliably extracted in the places that are unstructured and have no line segments. This drawback can be well compensated for by using the object features, and relatively slow recognition can be compensated for by using the line features in the environment where there are several line segments.

The second issue of this paper is localization. Localization is a method for estimating the pose of a robot with information from sensors mounted on the robot and an environmental map which is usually built by SLAM. Localization with low-cost

T. B. Kwon is with the Mechanical Engineering Department, Korea University, Seoul, Korea (e-mail: haptics@kora.ac.kr)

J. B. Song is with the Mechanical Engineering Department, Korea University, Seoul, Korea (Corresponding author: +82-2-3290-3363; e-mail: jbsong@korea.ac.kr)

M. Kim is with the Center for Intelligent Robotics, Korea Institute of Science and Technology, Seoul, Korea (e-mail: munsang@kist.re.kr).

sensors, however, 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 fusion of the range information from a low-cost IR scanner and the visual information from a stereo camera. The proposed localization scheme is mainly based on the MCL (Monte Carlo Localization) algorithm [6]. One problem involved in the fusion of the range and visual data is their different processing time. That is, the range information has a higher update rate than the visual information, because object recognition based on the SIFT feature extraction requires 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 with the encoder information. Then, dependable navigation is possible since the relatively poor range accuracy from the IR scanner can be compensated by the vision-based localization and the slow object recognition can be overcome by the frequent update of the range information.

The third issue is object detection. Some objects such as a desk or a chair are encountered in indoor navigation. A robot should avoid it or stop in front of it to interact with objects on it. It is very difficult, however, to perceive these objects in the real environment because a range sensor usually detects only desk legs and a vision sensor observes not only the desk surface but also other objects and background together. Therefore, a mobile robot can easily collide with these types of objects if the robot cannot detect it well.

This paper is focused on developing the simple and practically useful scheme to detect a desk. This method uses both range and vision sensors to exploit both sensors' advantages. By this approach, a robot can detect a desk while it moves in the real environment.

The remainder of this paper is organized as follows. Section 2 presents the SLAM scheme which uses both range data and recognized objects. Section 3 describes how to estimate the robot pose using both range data and recognized objects. Section 4 describes how to detect an object, especially a desk, using both range and vision sensors. Finally, section 5 presents conclusions.

## II. SLAM USING RANGE AND VISION SENSORS

### A. Feature-based SLAM

Several algorithms to extract lines from the measurement of range sensors have been studied [7]. Most algorithms are applied to laser scanners, but in this paper lines are extracted using an IR scanner by dividing range data into several line clusters and by applying a least square method to each line cluster [8]. An IR scanner provides a total of 121 range data

with a resolution of  $1.8^\circ$ . In case of the IR scanner, range data is limited to 2m, because the data exceeding 2m are found to be incorrect for line extraction. Figure 1(a) represents extracted line features using a least square method with the IR scanner.

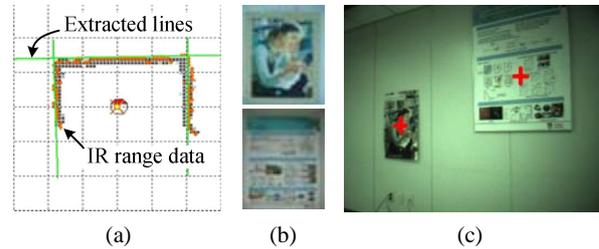


Fig. 1 Feature extraction; (a) line extraction using IR range data, (b) objects stored in DB, and (c) object recognition during navigation.

SIFT, one of the most popular recognition methods, is used to find an object registered in the database from the current image and to extract the relative range and angle from the robot to the object [9]. In this case, the center of the object is selected as a point representing this object. Figure 1(b) shows the objects registered in the database, and Fig. 1(c) depicts the recognized object during navigation.

In this research, SLAM is based on the extended Kalman filter (EKF), the most widely used filter in the SLAM of a mobile robot. It is suitable to fuse the range data from an IR scanner and the object information from a vision sensor. The EKF is based on the Bayes filtering and consists of the prediction and update stages. At the prediction stage, the states and uncertainties of a robot and features are predicted from the robot motion. At the update stage, the states and uncertainties are updated by the measurement of the sensor. The odometry error on the robot pose is compensated by the Kalman gain which is proportional to the difference between predictions and measurements. Figure 2 shows the basic concept of this approach.

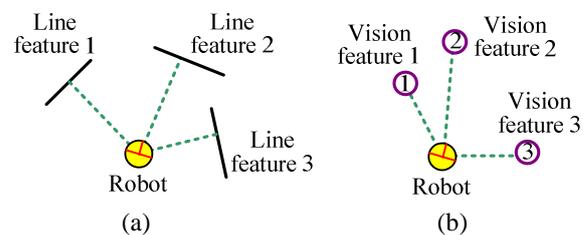


Fig. 2 Representation of features in the robot frames; (a) line features and (b) visual features.

### B. Experimental Results

The robot equipped with a Hokuyo PBS-03JN IR scanner and Videre STH-DCSG-C stereo camera was used for the experiments. Experiments were conducted in the environment of a living room, and the area was 9.5m x 7.5m in size with several pieces of furniture, as shown in Fig. 3(a). The

environment was modeled by a grid map at which each cell is 10cm x 10cm in size. Figure 3(b) shows 11 objects which were used as landmarks.



(a)



(b)

Fig. 3 Experimental environment and features; (a) experimental environment, and (b) objects used as visual features.

Figure 4 shows the processes of SLAM in the environment of Fig. 3. The information about the environment was not known, but the visual information of the objects was stored in the database in advance. In Fig. 4(b), only the object is recognized and in Fig. 4(d), only the lines are extracted. In Fig. 4(a) and (c), both the line extracted from the range data and the object recognized using the image are used together to localize the robot pose and map the environment simultaneously.

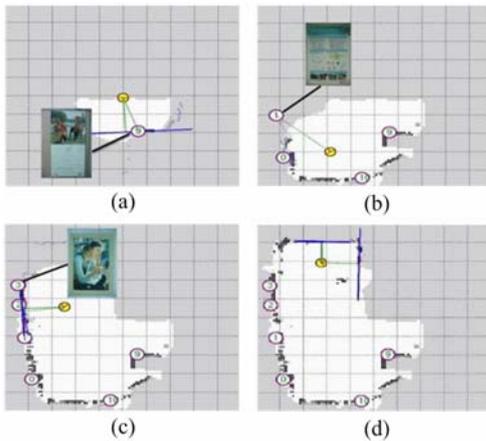


Fig. 4 Map building process of the proposed method.

Figure 5 is the map built by the proposed SLAM method. Each number means the object number which is shown in Fig. 3, and the position of the number means the position of that object on the map. Object 6, 7 and 8 are located beyond the line feature because the line feature is extracted from the table but the objects are hung on the wall. This result shows that the proposed method can work well in the environment which consists of both line features and visual features.

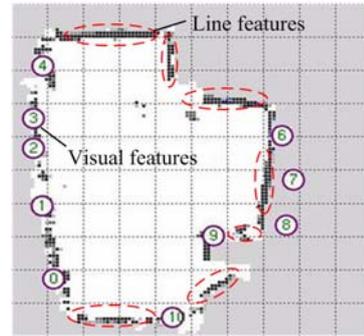


Fig. 5 Final grid map resulting from the proposed SLAM algorithm.

### III. LOCALIZATION USING RANGE AND VISION SENSORS

Range sensors such as a laser scanner and an IR scanner have been extensively used for global localization. However, when only range sensors are employed, the estimation error of the robot pose increases in dynamic or cluttered environments. It is also difficult to find the accurate robot pose when the robot is placed at a simple environment like a hallway. On the other hand, a vision sensor usually provides more information than a range sensor, and its performance is good for its cost. Therefore, substantial efforts have been directed toward the development of vision-based global localization for the past decade. In topological Markov localization [9], the input image was compared with the images stored at each node of the topological map and then the node at which the robot was located was found by Markov localization. A Vision-based SLAM was also proposed which used the SIFT (Scale Invariant feature transform) algorithm [10] based on the stereo vision [11].

In this research, the range and vision sensors are fused together for improved localization of a mobile robot. In this research, an IR scanner and stereo camera are used as main range and vision sensors respectively. 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 and Vision 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. 6(a). 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 as shown in Fig. 6(b). In Fig. 6, the measured range data to the environment,  $z_r$ , is obtained by the range sensor, and the predicted range data of  $i$ -th sample,  $z_r^{(i)}$ , is calculated based on the map. The relative angle to the recognized object from the robot at time  $t$ ,  $\theta_t^{obj}$ , is obtained on the image plane,

and the range from the robot to the recognized object,  $d_r^{obj}$ , is obtained by the stereo camera. The predicted angle and range to the object,  $\theta_r^{(i)}$  and  $d_r^{(i)}$ , are calculated based on the map likewise.

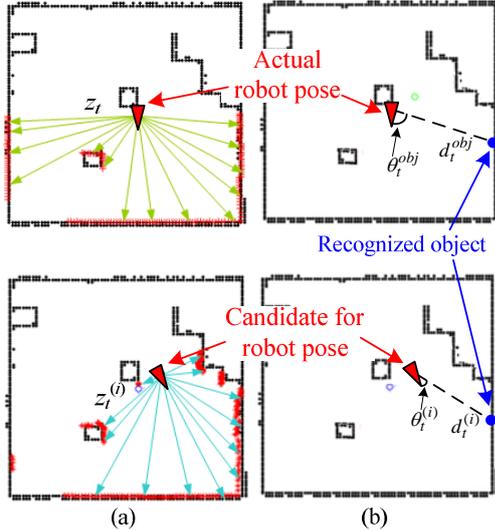


Fig. 6 Comparison of measured data and predicted data; (a) measured and predicted range data, and (b) observed and predicted object data.

### B. Fusion of Range and Vision Sensors

Vision-based localization can generally give more effective localization performance than range-based localization, provided objects with visual features abound 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 is used to update the probability of samples, and if the vision sensor recognizes any object, the range sensor model and the vision sensor model are fused to update the probability of samples.

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, the 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 the object recognition. For synchronization, therefore, the range data measured at the start of object recognition must be fused with the visual data at the end of object recognition. As the processing time for the object recognition increases, several sets of recent range data should be discarded for synchronization with the vision data. Thus, the overall update rate of the sample probability becomes low, thus leading to the increasing failure rate of localization due to lack of the most recent environment 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 samples while object recognition is in process. Sensor fusion is conducted only when the object recognition is completed and the visual data are available. Figure 7 shows the concept of this approach.

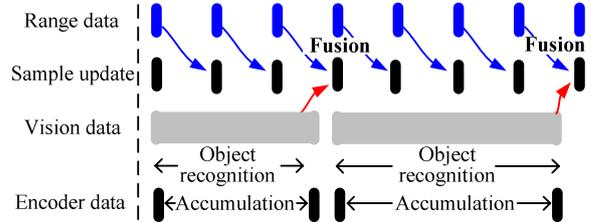


Fig. 7 Sensor fusion without loss of sensor information.

### C. Experimental Results

Various experiments have been performed with the robot equipped with an IR scanner (Hokuyo PBS-03JN) and a stereo camera (Videre STH-DCSG-C). As shown in Fig. 8(a), the experimental environment was 15m x 80m in size and consisted of a long hallway and several doors. The grid size of a grid map was 10cm. 11 visual landmarks were used as visual landmarks for localization. Figure 8(b) illustrates some of them, and their positions were shown as red dots in Fig. 8(a).

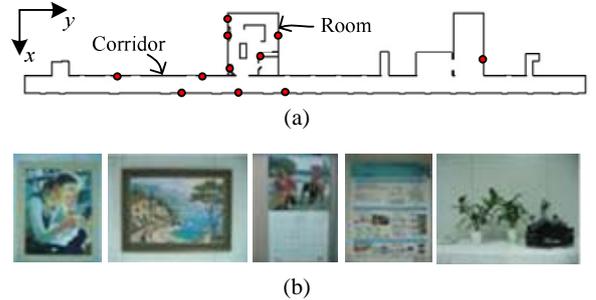


Fig. 8 (a) global map of experimental environment, and (b) objects used as visual landmarks.

Suppose the robot was placed at a room at the beginning of MCL. As shown in Fig. 9(a), sample variances converge to zero and thus the estimated robot pose can keep track of the actual one reasonably well because enough environmental information can be obtained only from 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 limit of the range sensor (4m in this experiment).

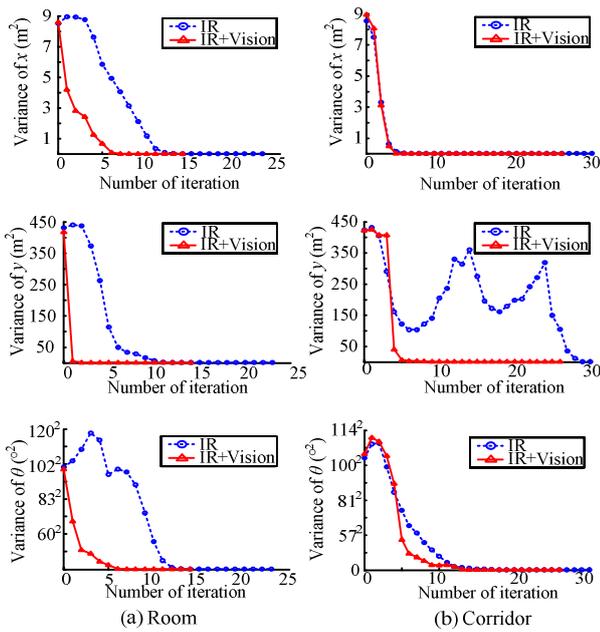


Fig. 9 Variance of sample position in  $x$ ,  $y$  axis and  $\theta$ ; (a) room, (b) hallway.

If a robot is located at a hallway at the beginning of MCL, the localization performance is generally worse than a robot being at 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. Even in this case, however, if the sensor fusion is conducted for localization, the better convergence of sample variances can be achieved, thus resulting in successful localization.

#### IV. OBJECT DETECTION USING RANGE AND VISION SENSORS

Several ways of detecting an object such as a desk have been studied so far. The method to detect a door and a table using line features of images were proposed [12]. In this research, the lines extracted from image were classified into a vertical line and a horizontal line, and compared with the models which were made by an operator. Another object detection method which uses the information about the color and 3D shape of the objects in its database was proposed in [13]. During navigation, a robot generated an image using the 3D information of objects, and compared that image with the image obtained by a vision sensor.

In this approach, the target object is a desk, which is one of the most important objects for navigation, and the range and vision sensors are used for reliable desk detection. A range sensor such as a laser scanner or an IR scanner detects the 2D environment, and in this case, only the desk legs can be sensed. Even though a range sensor can offer only 2D information, it can provide more accurate and stable data than a vision sensor. Therefore, good estimation of leg positions can be obtained using range sensor data.

After leg detection, the candidates for desk pose should be generated. In this case, the model information is required to generate the candidates which are similar to the real desk. Without the model, many candidates should be generated and evaluated to find the real table pose, but with the aid of the model information, the accurate and small candidates can be made. Then, all candidates should be evaluated to choose the best one. The surface of a desk is the most important part that has relatively plenty of features used to compare the candidates with the real desk. A vision sensor is suitable for performing this task. After calculating the similarity between each candidate and the vision sensor data, the best candidate is chosen and the desk pose is estimated.

##### A. Leg Detection Using Range Data

A leg of a common desk can be modeled as a pole. To detect a pole reliably using a range sensor is a very difficult task, especially for the inaccurate range sensor such as IR scanner. In this research, however, it is assumed that a laser scanner is used and the pole is reliably detected. The Hough transform for detection of circular arcs can be applied [14]. Figure 10 shows the result of detection, and not only the table legs but also human legs or cylindrical trash cans be detected. The numbers in Fig. 10 represent the distances (in mm) from the robot to the center of the detected poles.

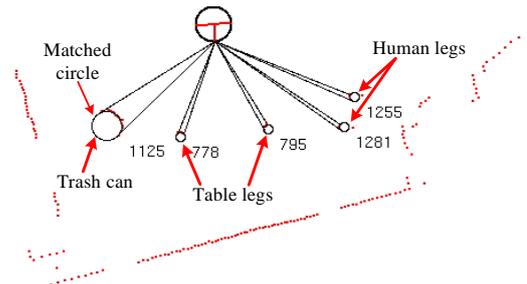


Fig. 10 Detection of arcs based on Hough transform.

##### B. Model Information

If a robot has no information on the desk, the whole area in which a desk can be located should be searched. In this case, it is difficult for a robot to detect a desk in real-time due to high computational complexity. To overcome this difficulty, the proposed method uses the minimum information on the desk, and figure 11 shows examples of the model information which was used in the experiments. Another advantage of the use of the model information is that a robot can detect a table although the whole table is not within the field of view (FOV) of a vision sensor.

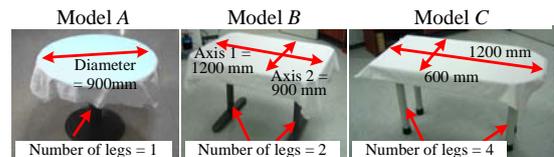


Fig. 11 Examples of model information.

### C. Candidate generation with detected legs

The candidates can be generated using both the model information and some poles selected among all detected poles. This task is executed by the candidate generator of each model. Model A with only one leg in Fig. 11 has a very simple candidate generator. It chooses one pole among all detected poles whose center positions are already found by the Hough transform. Then, it generates a candidate for every chosen pole. On the other hand, many combinations of two or more poles can be used to estimate the pose of Model C. It is possible that there exist other poles within the generated candidate because human legs or other objects can be detected as a pole. Table 1 shows the candidate generation rules of some models. It is easily obtained using geometry.

Table 1. Examples of candidate generation rules of some models.

	Model A No. of legs: 1	Model B No. of legs: 2	Model D No. of legs: 3	Model C No. of legs: 4
No. of considered poles: 1 		Too many possible candidates Meaningless	Too many possible candidates Meaningless	Too many possible candidates Meaningless
No. of considered poles: 2 	Already considered above			
No. of considered poles: 3 	Already considered above	Already considered above		
No. of considered poles: 4 	Already considered above	Already considered above	Already considered above	

● Detected poles    ◉ Predicted poles

### D. Vision-based Candidate Evaluation

Every candidate should be evaluated to determine how similar it is to the real table. In this research, the stereo vision based information is used to compare the candidates with the sensed environment. The simple and intuitive approach to calculating the reliability of the candidate is to compare the sizes of two areas; the size of the candidate and the area where the stereo vision information can be obtained on the inside of that candidate region.

First, the environment is divided into a discrete 2D grid of cells with uniform size, 100mm by 100mm in this research. The initial values of all cells are *empty*. Then the stereo vision based information is projected on the grid. The cells on which the stereo vision based information is projected change their values to *occupied*. In Fig. 12, the gray cells represent the occupied cells. A darker cell has more 3D points within its boundary than a lighter cell.

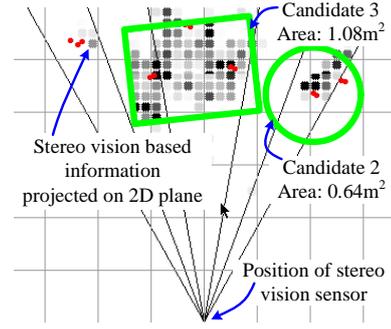


Fig. 12 Evaluation using candidate size and stereo vision based information.

These cells are used to evaluate the reliability of the candidates. If many cells within the area of one candidate have the occupied values, the reliability of that candidate is high as shown in candidate 3 in Fig. 12. The area of this candidate is 1.08m<sup>2</sup>, which corresponds to about 108 cells. The number of the occupied cells is 65. In this research, it is assumed that the candidate is reliable if more than half of all cells of the candidate are occupied. On the other hand, the area of candidate 2 is 0.64m<sup>2</sup> and it consists of about 64 cells. However, the number of the occupied cells within the area of this candidate is only 11, which means the reliability of this candidate is low. The reliability of a candidate is defined as

$$R_i = \frac{N_i}{A_i / 0.01} \quad (i = 1, \dots, n) \quad (1)$$

where  $n$  is the total number of all candidates,  $R_i$  is the reliability of the  $i$ th candidate,  $N_i$  is the number of the cells within the area of the  $i$ th candidate and  $A_i$  is the area of the  $i$ th candidate. After calculating the reliabilities of all candidates, the best candidate can be chosen to estimate the pose of a table.

### E. Experiments in Real Environment

By the proposed scheme, the tables on which a robot has information can be detected in the real environment. Figure 13 shows the experimental results. Only candidates for model B are drawn in Fig. 13 because the image would be very unclear if all candidates were drawn together. In Fig. 13(a), only two legs are detected and one candidate for model B is generated and its reliability is computed as 0.50. This candidate is chosen as the pose of a table and is projected on the right image of Fig. 13(a). In Fig. 13(b) and (c), many candidates are generated because one pole and a chair are added and human legs are also detected. From (a) to (c), however, one candidate whose pose is similar to the pose of a real table is chosen. The value of the reliability varies little because the stereo vision based information continues to change little.

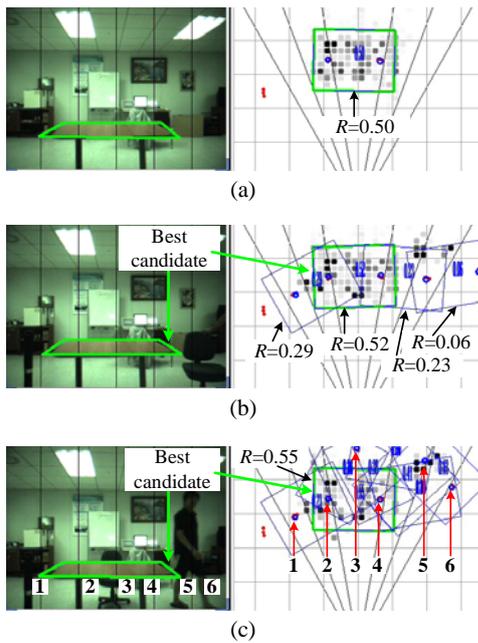


Fig. 13 Evaluation using candidate size and stereo vision based information.

## V. CONCLUSIONS

This paper proposed the practical approaches to the three important abilities for mobile robot navigation in the indoor environment. They use both the range and vision sensors to compensate for the drawbacks of each sensor. By this approach, more reliable and practical performance can be obtained.

From these researches, the following conclusions have been drawn.

1. Two types of features extracted using the range and vision sensors respectively are used together for the observations of the EKF, and it makes the SLAM more stable and practical than the case in which only one type of feature is used.

2. Sensor fusion based localization proposed here enables samples in MCL to converge to the actual robot pose faster than either range-based or vision-based localization alone. Although the processing time for object recognition takes a long time and is not periodic, the probability of samples can be updated at a speed of a range sensor with the proposed method.

3. In the real environment, many candidates for the desk pose should be generated with all detected poles. In order to choose the best one among many candidates, a comparison between the stereo vision based information and each candidate for the table pose in 3D space can serve as a good evaluation tool.

## ACKNOWLEDGMENT

This research was conducted by 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.

## REFERENCES

- [1] M. Montemerlo and S. Thrun, "Simultaneous Localization and Mapping with Unknown Data Association Using FastSLAM," *Proc. of IEEE Int. Conf. on Robotics and Automation*, pp. 1985~1991, 2003.
- [2] J. E. Guivant and E. Nebot, "Optimization of the Simultaneous Localization and Map-Building Algorithm for Real-Time Implementation," *IEEE Tran. on Rob. and Auto.*, Vol. 17, No. 3, pp. 242~256, 2001.
- [3] G. Grisetti, C. Stachniss and W. Burgard, "Improving Grid-based SLAM with Rao-Blackwellized Particle Filters by Adaptive Proposals and Selective Resampling," *Proc. of the 2005 IEEE Int. Conf. on Rob. and Auto.*, pp. 2443~2448, 2005.
- [4] D.G. Lowe, "Distinctive image features from scale invariant keypoints," *Int. Jour. of Comp. Vis.*, Vol. 60, No. 2, pp. 91~110, 2004.
- [5] S. Se, D.G. Lowe and J. J. Little, "Vision-Based global localization and mapping for mobile robots," *Trans. on Rob.*, Vol. 21, No. 3, pp. 364~375, 2005.
- [6] T.-B. Kwon, J.-H. Yang, J.-B. Song, W. Chung, "Efficiency Improvement in Monte Carlo Localization through Topological Information," *Proc. of Int. Conf. on Intell. Rob. and Sys.*, Oct. 2006.
- [7] V. Nguyen, A. Martinelli, N. Tomatis and R. Siegwart, "A comparison of Line Extraction Algorithms using 2D Laser Rangefinder for Indoor Mobile Robotics," *Proc. of Int. Conf. on Intell. Rob. and Sys.*, pp.1864~1869, 2005.
- [8] L. Zhang, "Line Segment Based Map Building and Localization Using 2D Laser Rangefinder," *Proc. of Int. Conf. on Rob. and Auto.*, pp. 2538~2543, 2000.
- [9] J. Kosecka and F. Li, "Vision based topological Markov localization," *Proc. of Int. Conf. on Rob. and Auto.*, vol. 2, pp. 1481-1486, 2004.
- [10] D.G. Lowe, "Distinctive image features from scale invariant keypoints," *Int. Jour. of Comp. Vis.*, vol. 60, no 2, pp. 91-110, 2004.
- [11] D.G. Lowe and S. Se, "Vision-Based global localization and mapping for mobile robots," *Trans. on Rob.*, vol. 21, pp. 217-226, June, 2005.
- [12] D. S. Kim and R. Nevatia, "A Method for Recognition and Localization of Generic Objects for Indoor Navigation," *Proc. of Workshop on Applications of Computer Vision*, Vol. 2, pp. 1069-1076, 1994.
- [13] H. Takeda, A. Ueno, M. Saji, T. Nakano and K. Miyamoto, "A Robot Recognizing Everyday Objects," *Proc. of Int. Conf. on Intell. Rob. and Sys.*, Vol. 2, pp. 1107-1112, 2000.
- [14] P. Kierkegaard, "A Method for Detection of Circular Arcs Based on the Hough Transform," *Trans. on Machine Vision and Applications*, Vol. 5, No. 4, pp 249-263, 1992.