Proceeding of the 2006 IEEE
International Conference on Automation Science and Engineering
Shanghai, China, October 7-10, 2006

Vision-based Indoor Scene Cognition Using a Spatial Probabilistic Modeling Method

Jwu-Sheng Hu[1], Tzung-Min Su[2], Heng-Chia Huang[3] and Pei-Ching Lin[4]
Dept. of Electrical and Control Engineering
National Chiao-Tung University
Hsinchu, Taiwan, China
[1]*jshu@cn.nctu.edu.tw;* [2]*linux.ece89g@nctu.edu.tw;* [3]*jans.ece93g@nctu.edu.tw;* [4]*peiching.ece93g@nctu.edu.tw*

***Abstract* – This work describes a vision-based approach to recognize scene in the indoor environment. The proposed method represents each scene captured by a Pan-Tilt-Zoom (PTZ) camera with a blob model using spatial probabilistic modeling. Although the details of the scene covered by the camera are lost, this model is efficient in memorizing the scene characteristics and is robust against image distortions. Furthermore, multi-view recognition is studied to increase the precision of scene cognition via a partial knowledge of the scene. The images captured in the same location with different view angles are collected to extract the scene characteristics in order to decrease the memory storage size for each location. The effectiveness of the method is demonstrated by experiments in an unstructured indoor environment.**

***Index Terms* – characteristic view, Gaussian mixture model, probabilistic modeling, scene cognition**

I. INTRODUCTION

Scene cognition is a fundamental problem in mobile robot localization. It is one of the common ability of human but is difficult in the field of computer vision. In the past, many researches for mobile robot localization have been studied and they mainly differ in the internal representation of the environment. First, a geometrical representation of space is adopted to provide the path to be followed by a robot via different degrees of detail, varying from a complete CAD model of the environment to a simple graph of interconnections or interrelationships between the elements in the environment [1][2]. These geometrical approaches are based on either map matching or landmark detection. Most map matching systems rely on a good estimation of the robot location and are not suitable in the populated environment where the robot may collide with humans or other obstacles. As for the landmark localization systems, either artificial or natural landmarks are utilized to localize the robot location. However, they are not easy to be applied to different environments. Second, a topological representation of space is utilized to describe the environment as an adjacency graph, where the node of the graph corresponds to the robot's location [3][4]. Although topological maps are less accurate than geometrical maps, they are also less complex and easier to be generated and maintained than geometrical maps.

In order to provide rich information to distinguish adjacent locations, the color vision camera is often adopted to build up the topological maps. The PTZ camera and omni-directional camera are two alternates used in capturing scenes. Although the omni-directional camera has a 360^{o} view angle, it suffers from image distortion and occlusion. For PTZ camera, the view angle limitation decreases the covered area but the ability of pan and tilt allows the robot to capture multi scenes around a position to overcome the influences caused from occlusion. However, the dimension of the incoming sensor information from either omni-directional camera or PTZ camera is very high and needs enormous memory storage.

Principal component analysis (PCA) is a general approach to reduce the dimensionality of the data space by describing the captured image with a feature vector that containing only a limited number of linear PCA features [5][6]. In this work, a probabilistic spatial model of the captured image, called a blob model, is proposed to reduce the memory storage of scenes. Two approaches were taken in modeling the probabilistic representation: the first one is called parametric method, which uses single Gaussian distribution [7] or mixtures of Gaussian [8][9]; the second one is called non-parametric method, which uses the kernel function to estimate the density function [10][11]. Furthermore, the concept of characteristic view, which was first presented in the field of object recognition [12], was adopted to train the scene model of a position. It is convenient for the mobile robot to train and update the scene model.

The novelty of this work is using the blob-model to represent the scene in the indoor environment and combining the blob models around a position to a compact set of scene model. It can decrease the memory storage of scene model and improve the flexibility and robustness of training the nodes of the topological maps. The rest of this paper is organized as follows. Section II describes the features and the statistical learning method used in the spatial probabilistic modeling. In Section III, the scene combinational algorithm is described and the value of the likelihood function calculated using the features from the captured image is used to recognize scene in the indoor environment. Subsequently, experimental results are shown to demonstrate the performance of the proposed

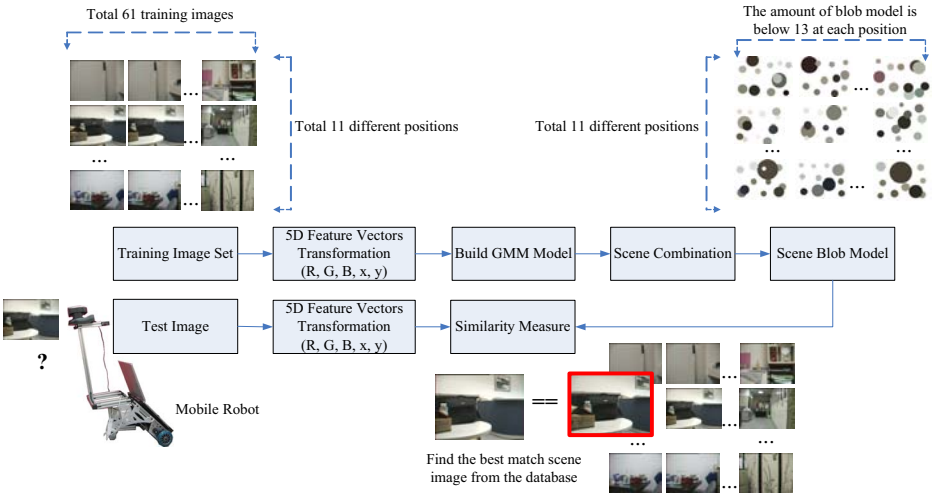


Fig. 1 Basic workflow of the proposed framework

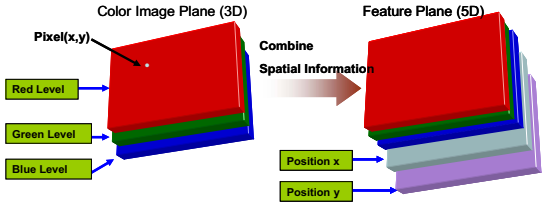


Fig. 2 5D feature vector construction

method in scene cognition. A conclusion of this work is in Section V. Moreover, the basic workflow of the proposed framework is illustrated in Fig. 1.

II. A Probabilistic Scene Model in Spatial Domain

While humans have no difficulty in extracting conceptual information regarding an image, computers have difficulties in doing the same thing. Methods for conceptual level background modeling were used to derive the image representation, such as histogram [13], the Gaussian Mixture Model (GMM), Kernel Density Estimation, and so on. Moreover, feature selection is a key source of difference between different applications using these methods.

A. Feature Extraction

Many features have been utilized to extract useful information from the captured image, such as edge, corner, texture, color and shape. Among these features, color involves the intuitive information to represent the conceptual idea of an image. Therefore, pixel color and pixel position are utilized in this work to extract the conceptual idea of an image. The color space used here is the RGB color space, which is common for most video devices. To enhance the regional information of an image, the position (x, y) feature is combined with RGB color information to be the feature vector. That is, each pixel contains a 5D feature vector (R, G, B, x, y), which is shown in Fig. 2.

B. Modeling in the Spatial Domain

This work applies GMM to model the region information in an image using the 5D feature vectors (R, G, B, x, y). It is assumed that the density function of the color and position features both have Gaussian distributions. First, each pixel x is defined as a d-dimensional vector at time t (In this work, d is defined as 5). N Gaussian distributions are used to construct the GMM, which is described as follows:

$$f(x \mid \lambda) = \sum_{i=1}^{N} w_i \frac{1}{\sqrt{(2\pi)^d \mid \Sigma_i \mid}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right) \quad (1)$$

λ represents the parameters of GMM,

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1,2,...,N \text{ and } \sum_{i=1}^{N} w_i = 1$$

Next, the parameters λ of GMM are calculated to enable the GMM to match the feature vector distribution with the least error. The most widespread method is the maximum likelihood (ML) estimation. The objective of ML estimation is to identify model parameters by maximizing the likelihood function of GMM obtained from the training feature vectors X. ML parameters can be derived iteratively using the expectation maximization (EM) algorithm [14]. Supposing there are m feature vectors $x_1, x_2, ..., x_m$ (In this work, m is defined as the image size, 320x240=76800), then the maximum likelihood estimation of λ can be calculated via equation (2).

$$\lambda_{ML} = \arg\max_{\lambda} \sum_{j=1}^{m} \log f(x_j \mid \lambda) \quad (2)$$

The EM algorithm involves two steps; the parameters of GMM can be obtained by iteratively computing the following equations (equations (3)-(4)):

- Expectation step: (E step)

$$\beta_{ji} = \frac{w_i f(x_j \mid \mu_i, \Sigma_i)}{\sum_{k=1}^{N} a_k f(x_j \mid \mu_k, \Sigma_k)}, i = 1,...,N, \text{j} = 1,...,m \quad (3)$$

β_{ji} means the posterior probability that the feature vector x_j belongs to the *ith* Gaussian component distribution.

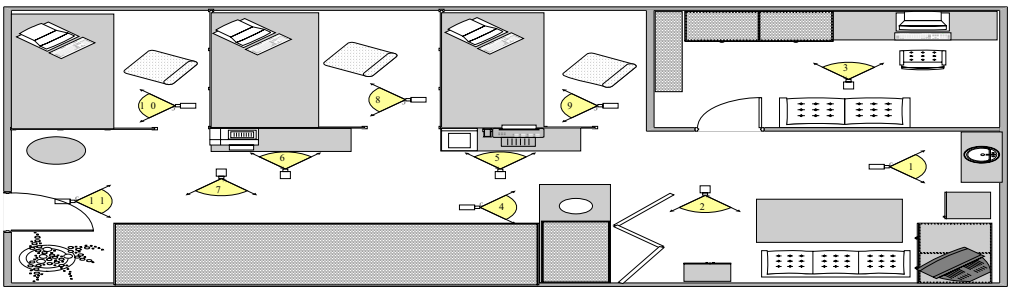


Fig. 3 The indoor environment from which images were taken

- Maximum step: (M step)

$$\hat{w}_i = \frac{1}{N}\sum_{j=1}^{m}\beta_{ji}$$
$$\hat{\mu}_i = \sum_{j=1}^{m}\beta_{ji}x_j \Big/ \sum_{j=1}^{m}\beta_{ji} \tag{4}$$
$$\hat{\Sigma}_i = \sum_{j=1}^{m}\beta_{ji}(x_j-\hat{\mu}_i)(x_j-\hat{\mu}_i)^T \Big/ \sum_{j=1}^{m}\beta_{ji}$$

The termination criteria of the EM algorithm are as follows:

a. The increment between the new log-likelihood value and the last log-likelihood value is below a minimum increment threshold.
b. The iterative count exceeds a maximum iterative count threshold.

Unsupervised data clustering is used before the EM algorithm iterations to accelerate the convergence. This study uses the K-means algorithm [15] for the clustering. The number of cluster is defined and then the initial center of each cluster is selected randomly. The suitable center and variance of each cluster can be estimated iteratively via the K-means algorithm and applied to be the initial mean and variance of each Gaussian component of GMM.

III. Scene Representation and Cognition

Although the omni-directional camera has a 360° view angle, it suffers from image distortion and occlusion. In this work, PTZ camera is adopted to capture multi-directional image at a position to overcome the influences caused from occlusion. First, different directional images are captured at one position to represent the scene of a position in the training phase. Besides, multi-views can be utilized to recognize the scene. However, the memory storage of these scene models is enormous. In this section, the concept of characteristic view, which was first presented in the field of object recognition [12], was adopted to extract the compact set of the scene model. Besides, it is convenient for the mobile to train and update the scene model.

A. Scene Representation via Characteristic views

While a large number of 2D views are collected by the mobile robot, the scene can be described more detailed, but the computing time to recognize the scene is consequently growing due to huge searching space. Therefore some methods are studied to extract a minimal set of object views. Extracting the characteristic views from these 2D views is a kind of approach to obtain a compact set of scene views. In our previous work [16], three kinds of similarity measures, 1-norm, 2-norm and K-L distance, have been applied to extract the characteristic views with the proposed combinational algorithm. In this work, K-L distance is utilized to calculate the similarity between blob-models.

Suppose p_0, p_1 are two probability densities, the K-L distance is defined as

$$D'(p_1 \| p_0) \approx \sum_{t=1}^{L}\left(p_1(x_t)\cdot\log\left(\frac{p_1(x_t)}{m(x_t)}\right) + p_0(x_t)\cdot\log\left(\frac{p_0(x_t)}{m(x_t)}\right)\right)$$
where $m(x_i) = (p_0(x_i)+p_1(x_i))/2$ (5)

In this work, p_0 and p_1 can be calculated from the blob model of each image via their Gaussian mixture model, which is described as equation (1).

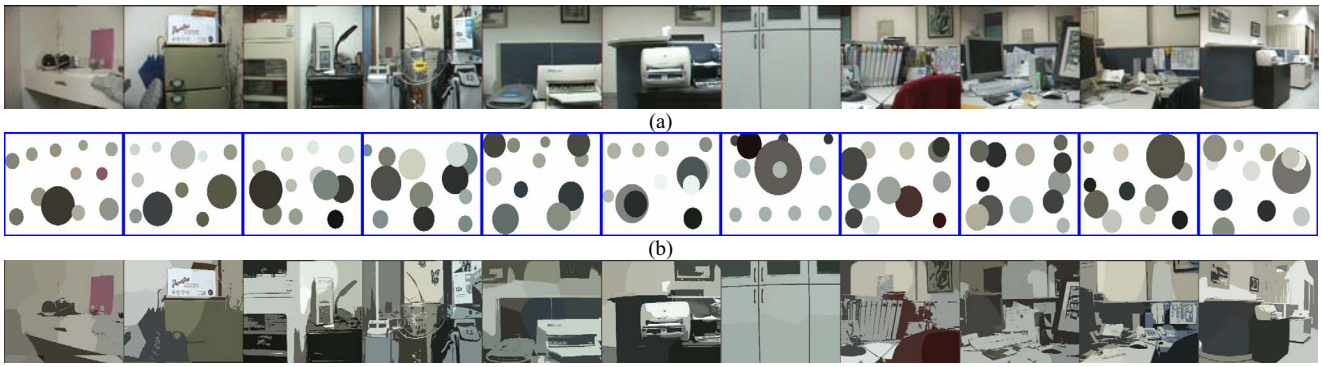
Suppose N_c 2D views are captured at a position (the pan angle of the PTZ camera is between $-\theta_p$ and θ_p, and the tilt angle of the PTZ camera is zero, but it is easily extended to include tilt motion) to train the scene blob model. Then N_r ($N_r \le N_c$) 2D views are extracted as the set of scene model using a combinational algorithm described in our previous work [16]. If S positions in the environment are selected as the nodes of topological maps, the amount of overall blob models stored in the database can be described as N_t, where

$$N_t = \sum_{i=1}^{S} N_r(i) \tag{6}$$

where $N_r(i)$ denotes the amount of blob-model at the i_{th} position.

B. Multi-view Scene Cognition via ML Values

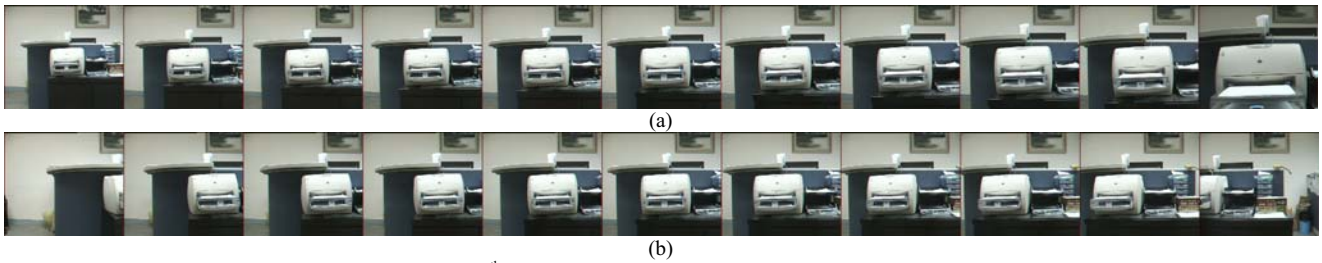
In order to increase the robustness of the result of scene cognition, multi-views are captured at a location by the PTZ camera. Three arbitrary images I_i ($1 \le i \le 3$) of different pan and tilt angles of the PTZ camera are utilized to calculate the Maximum Likelihood values of the N_t blob models. The first three recognized results, having the first three minimum Maximum Likelihood values, estimated by comparing the test image and the database are selected as the candidates to be further processed.



(a) (b) (c)

Fig. 4. The sample of the training image, blob model and conceptual description of each scene captured in the indoor environment (Fig. 3), (a) The sample of captured image at each location in the indoor environment (from left to right is the position 1,2,…,11), (b) The blob model of each sample of captured image in (a) with 12 Gaussian distribution, (c) The conceptual description of each sample of captured image in (a), which are calculated by comparing the original pixel values of each captured image with its blob model.

Fig. 5. The 11 characteristic views at the 5th position in the indoor environment.



(a) (b)

Fig. 6. The sample of test images captured from the 5th position in the indoor environment, (a) the test images captured with forward and backward direction, and the shifted distance is backward 50cm, backward 20cm, backward 15cm, backward 10cm, backward 5cm, 0cm, forward 5cm, forward 10cm, forward 15cm, forward 20cm and forward 50cm respectively, (b) the test images captured with left and right direction, and the shifted distance is left 50cm, left 20cm, left 15cm, left 10cm, left 5cm, 0cm, right 5cm, right 10cm, right 15cm, right 20cm and right 50cm respectively.

Suppose V denotes the set of recognized result, then V is defined as follows:

$$V = \{v_{ij}\}, 1 \le i \le 3, 1 \le j \le 3, 1 \le v_{ij} \le S,$$

where

i: the index of test image

j: the index of the order of recognition result

Moreover, three methods are proposed here to estimate the final result of scene cognition. The first result R_1 is estimated only using one captured image and the first minimum ML value. The second result R_2 also uses one captured image but all the first three minimum ML values are utilized to improve the robustness of the recognition result. The third result R_3 uses all the three captured images and all their first three minimum ML values. The descriptions of R_1, R_2 and R_3 are described as equations (7)-(10):

$$R_k = \begin{cases} v_{11} & , k = 1 \\ v_{11} \cdot \bar{D}_1 + m_1 \cdot D_1 & , k = 2 \\ v_{11} \cdot (\bar{D}_1\, \bar{D}_2\, \bar{D}_3) + m_1 \cdot (D_1) + \bar{D}_1[m_2 \cdot (D_2) + \bar{D}_2(m_3 \cdot D_3)], & k = 3 \end{cases} \quad (7)$$

Fig. 7. The mobile robot ER1

where

$$m_p = \arg\max(F_p) \quad , 1 \le p \le 3 \quad (8)$$

$$D_p = \begin{cases} 1, if\ \text{argmax}(F_p)\ \text{exists} \\ 0, if\ \text{argmax}(F_p)\ \text{doesn't exist} \end{cases}, 1 \le p \le 3 \quad (9)$$

$$F_p = \{f_{pq}, 1 \le q \le 11\}, \ f_{pq} = \sum_{j=1}^{3} \delta(q - v_{pj}) \quad (10)$$

IV. Experimental Results

This section describes two experiments that demonstrate the effectiveness of the proposed method. Real image sequences are acquired with the Sony EVI-D30 PTZ camera that mounted on the mobile robot ER1.

TABLE I
THE ROBUST TEST VIA POSITION VARIATIONS

Shift Distance and Direction (The number of test image)		Recognition Rate (1.0000=100%)		
		R_1	R_2	R_3
0 cm (61 pictures, self-test)		1.0000	1.0000	1.0000
5 cm	Forward (61 pictures)	1.0000	1.0000	1.0000
	Backward (61 pictures)	1.0000	0.9985	1.0000
	Left (61 pictures)	1.0000	1.0000	1.0000
	Right (61 pictures)	1.0000	1.0000	1.0000
10 cm	Forward (61 pictures)	1.0000	1.0000	1.0000
	Backward (61 pictures)	1.0000	0.9985	1.0000
	Left (61 pictures)	1.0000	0.9911	1.0000
	Right (61 pictures)	1.0000	1.0000	1.0000
15 cm	Forward (61 pictures)	1.0000	1.0000	1.0000
	Backward (61 pictures)	1.0000	0.9940	1.0000
	Left (61 pictures)	0.9955	0.9777	1.0000
	Right (61 pictures)	0.9918	0.9787	0.9967
20 cm	Forward (61 pictures)	1.0000	1.0000	1.0000
	Backward (61 pictures)	1.0000	0.9926	1.0000
	Left (61 pictures)	0.9747	0.9553	0.9821
	Right (60 pictures)	0.9762	0.9568	0.9836
50 cm	Forward (61 pictures)	0.8495	0.8495	0.8554
	Backward (61 pictures)	0.8972	0.8540	0.9344
	Left (61 pictures)	0.7526	0.7481	0.7735
	Right (61 pictures)	0.8271	0.7973	0.8525

Fig. 7 illustrates the ER1 robot equipped with the PTZ camera. The size of the captured image is 320 by 240. For demonstration, only the images captured with the pan motion of the PTZ camera are utilized to extract the scene model. However, the proposed method can also be applied to a combined motion via more elaborate calculation. Training images of 11(S=11) locations in the environment (Fig. 3) are obtained by rotating the camera from -30 to 30 degrees using 1 degree increments at each location. Therefore, 61 images are captured at each position and N_c is then defined as 61. In this work, each scene model is established with 12 Gaussian distributions (N=12) with the 5D feature vectors described in section II. Besides, the characteristic views of the scene model at each position are extracted using the combinational algorithm in our previous work [16] and the amount of the characteristic views at each position are all below 13 after combination ($N_r(i) \le 13, 1 \le i \le 11$). The sample of the training images, blob models and conceptual descriptions of each scene captured in the indoor environment (Fig. 3) are listed in Fig. 4. Besides, for the sake of illustration, the set of the characteristic views at the 5th position in the indoor environment is cited as an instance and is listed in Fig. 5. The experiments are segmented into 2 stages, described as below:

A. Experiments for robustness via position variations

In the first experiment, a self-test is first performed by testing the 61 training images with the extracted scene model N_t. Then the mobile robot moves in four directions (forward, backward, left and right) with five different distances (5cm, 10cm, 15cm, 20cm and 50cm). At each position, 61 test images are captured by rotating the camera from -30 to 30 degrees using 1 degree increments with no occlusion in the scene. The samples of test images captured at the 5th position are listed in Fig. 6, where Fig. 6(a) shows the test images captured with forward and backward direction and Fig. 6(b) shows the test images captured with left and right direction. In Table I, the three kinds of results, R_1, R_2 and R_3, are all listed to evaluate the performance.

From the results listed in the Table I, the recognition rates of the three kinds of method are all above 95% when the position variations are below 20cm. Besides, the third method R_3 has the best performance than others. It makes sense to base the perceptual skills used for localization on vision, like humans do. When a person comes into an unknown place, multi-directional views are captured by the eyes to help to recall the memory of his/her past experience about the unknown place. In this work, the same strategy is adopted to increase the robustness of scene cognition.

B. Experiments for robustness via position variations and different level of occlusion

During the second stage, not only position variations but also different levels of occlusion are considered in this experiment. The position variations are the same as those in the first experiment, 4 different directions with 5 different distances shifted from the original S positions. Different obstacles are put into the covered area of the PTZ camera to simulate the case of occlusion, where 5%, 10%, 15%, 20% and 50% regions are different from the original test image captured in the first experiment. In Table II, the recognition rates are also above 95% when the level of occlusion is less than 20% and the position variations are below 20cm. Even though the level of occlusion arrives 50% and the position variations are 50cm, the recognition rates are still above 50%. Besides, the third method R_3 still has the best performance than others.

V. CONCLUSIONS

This study proposes a vision-based approach to recognize scene in the indoor unstructed environment. The effectiveness of the method is demonstrated by experiments in an unstructured indoor environment. The robustness of the scene blob model is performed via recognizing those test images with different levels of position variations and occlusion. Furthermore, the memory storage of the N_c training images is decreased by replacing the original color images with only the parameters of the Gaussian mixture model and removing the redundant training images using the concept of characteristic view, which was first presented in the field of object recognition. Although the details of the scene covered by the camera are lost, this characteristic blob-model is efficient in memorizing the scene characteristics. Moreover, multi-view recognition strategy is applied to increase the robustness of scene cognition.

These advantages permit the proposed method to be used in situations with limited memory size and in the

TABLE II
THE ROBUST TEST VIA POSITION VARIATIONS AND DIFFERENT LEVEL OF OCCLUSION

Shift Distance (cm) and Direction		Covering Rate (1.000=100%)														
		5%			10%			15%			20%			50%		
		R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3
0 cm		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.800	0.769	0.817
5	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.797	0.775	0.809
	Backward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	0.794	0.763	0.809
	Left	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.794	0.779	0.806
	Right	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.791	0.754	0.818
10	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.796	0.784	0.802
	Backward	0.997	0.979	1.000	1.000	0.997	1.000	1.000	0.997	1.000	0.997	0.996	1.000	0.785	0.738	0.802
	Left	1.000	0.993	1.000	1.000	0.996	1.000	1.000	0.996	1.000	1.000	0.993	1.000	0.796	0.772	0.805
	Right	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000	0.999	1.000	0.770	0.747	0.817
15	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.784	0.769	0.797
	Backward	1.000	0.996	1.000	1.000	0.993	1.000	0.997	0.988	1.000	0.994	0.987	1.000	0.763	0.726	0.781
	Left	0.997	0.979	1.000	0.997	0.979	1.000	0.997	0.979	1.000	0.994	0.975	1.000	0.779	0.736	0.794
	Right	0.992	0.982	0.997	0.992	0.977	0.997	0.992	0.979	0.997	0.992	0.977	0.997	0.726	0.705	0.757
20	Forward	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.765	0.751	0.782
	Backward	0.999	0.987	1.000	0.994	0.982	1.000	0.991	0.979	1.000	0.988	0.976	1.000	0.733	0.711	0.748
	Left	0.976	0.960	0.987	0.979	0.970	0.993	0.975	0.961	0.979	0.975	0.963	0.979	0.748	0.699	0.776
	Right	0.975	0.955	0.984	0.979	0.970	0.993	0.976	0.951	0.984	0.975	0.951	0.984	0.714	0.694	0.744
50	Forward	0.845	0.838	0.854	0.845	0.844	0.849	0.842	0.829	0.845	0.832	0.815	0.839	0.508	0.503	0.523
	Backward	0.881	0.839	0.925	0.848	0.821	0.896	0.830	0.809	0.872	0.796	0.785	0.833	0.525	0.508	0.553
	Left	0.750	0.741	0.775	0.748	0.742	0.768	0.733	0.729	0.754	0.723	0.711	0.735	0.502	0.501	0.531
	Right	0.811	0.794	0.841	0.799	0.784	0.827	0.781	0.770	0.817	0.753	0.763	0.778	0.532	0.502	0.531

populated or unstructed environment. Moreover, the computation requirement for the proposed method remains high especially when the number of Gaussian component used for the spatial probabilistic modeling increases. In this work, the feature vector of each pixel is adopted to calculated the similarity between the test view and views in the database in this work. Therefore, selecting the noticeable pixels is one way to reduce the computing time. However, the real-time issue is still the future work of this study.

ACKNOWLEDGMENT

This work was supported by National Science Council of the R.O.C. under grant no. NSC93－2218－E009064.

REFERENCES

[1] G.N. Desouza, A.C. Kak, "Vision for mobile robot navigation: a survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 237-267, 2002.

[2] A. Kosaka and A.C. Kak, "Fast vision-guided mobile robot navigation using model-based reasoning and prediction of uncertainties," Computer Vision, Graphic, and Image Processing -- Image Understanding, Vol. 56, No. 3, November, pp.271-329, 1992.

[3] I. Ulrich and I. Nourbakhsh, "Appearance based place recognition. for topological localization," *in IEEE Conf. on Robotics and. Automation*, November 2000, pp. 1023–1029.

[4] P. Lamon, A. Tapus, E. Glauser, N. Tomatis and R. Siegwart, "Environmental modeling with fingerprint sequences for topological global localization," *in IEEE International Conf. on Intelligent Robots and Systems*, October 2003, pp. 3781–3786.

[5] B.J.A. Krose, N. Vlassis, R. Bunschoten, Y. Motomura, "A Probabilistic Model for Appearance-Based Robot Localization," *Image and Vision Computing, volume 19, pages 381--391, 2001.*

[6] B.J.A. Krose, R. Bunschoten, "Probabilistic Localization By Appearance Models and Active Vision," *In Proc. of the IEEE International Conf. on Robotics and Automation*, pp. 2255-2260, 1999.

[7] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 780–785, July 1997.

[8] C. Stauffer, and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking" in *Proc. CVPR*, v.2, pp.246-252, June 1999.

[9] P. KaewTrakulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Proc. 2nd Europcan Workshop on Advance Video Based Surveillance Systems*, AVBS01, Sept. 2001.

[10] A. Elgammal, D. Harwood, L. Davis, "Non-parametric Model for Background Subtraction," in *6th European Conf. on Computer Vision.* Dublin, Ireland, June/July 2000.

[11] A. Elgammal, R. Duraiswami, D. Harwood, L.S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," in *Proc. of the IEEE*, vol. 90, July 2002, pp.1151-1163.

[12] C. M. Cyr and B. Kimia, "A Similarity-Based Aspect-Graph Approach to 3D Object Recognition," in International Journal of Computer Vision, 57(1):5–22, 2004.

[13] N. Vasconcelos, A. Lippman, "Feature representations for image retrieval: beyond the color histogram," in *Proc. of the International Conf. on Multimedia and Expo*, New York, 2000.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Roy. Stat. Soc., vol. 39, no. 1, pp. 1-38, 1977.

[15] J. B. MacQueen, **"**Some Methods for classification and Analysis of Multivariate Observations," in *Proc. of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, vol. 1, pp. 281-297, 1967.

[16] T.M. Su, C.C. Lin, P.C. Lin and J.S. Hu, "Shape Memorization and Recognition of 3D Objects Using a Similarity-Based Aspect-Graph Approach," *In Proc. of the IEEE International Conf. on Systems, Man and Cybernetics,* October, 2006. (to be submitted).