Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 149206, 10 pages
http://dx.doi.org/10.1155/2015/149206

Research Article

Vision-Based Semantic Unscented FastSLAM for Indoor Service Robot

Xiaorui Zhu,[1] Fucheng Deng,[1] Yongsheng Ou,[2] Letian Liu,[1] and Ermeng Wang[1]

[1]*University Town of Shenzhen, HIT Campus, Nanshan District, Shenzhen 518055, China*
[2]*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*

Correspondence should be addressed to Xiaorui Zhu; hit.zhu.xr@gmail.com

Received 19 December 2014; Accepted 8 June 2015

Academic Editor: Shaoping Bai

This paper proposes a vision-based Semantic Unscented FastSLAM (UFastSLAM) algorithm for mobile service robot combining the semantic relationship and the Unscented FastSLAM. The landmark positions and the semantic relationships among landmarks are detected by a binocular vision. Then the semantic observation model can be created by transforming the semantic relationships into the semantic metric map. Semantic Unscented FastSLAM can be used to update the locations of the landmarks and robot pose even when the encoder inherits large cumulative errors that may not be corrected by the loop closure detection of the vision system. Experiments have been carried out to demonstrate that the Semantic Unscented FastSLAM algorithm can achieve much better performance in indoor autonomous surveillance than Unscented FastSLAM.

1. Introduction

Visual simultaneous localization and mapping (SLAM) uses the cameras as the only exteroceptive sensors to recover a representation of the environment and achieve localization of the robot complemented with information from the proprioceptive sensors with the aim of increasing accuracy and robustness. To the mobile robotic, vision has proved to be an effective and inexpensive sensing device for localization and mapping. Sim et al. solved the SLAM problem with a stereo pair of cameras [1, 2]. Schleicher et al. used a top-down Bayesian method to perform a vision-based mapping process where identification and localization of the natural landmarks from the images were provided by a wide-angle stereo camera [3]. In this paper, a new semantic vision SLAM framework is proposed to improve the performance without increasing the complexity of the algorithms dramatically.

Literature on visual SLAM have focused on feature-based SLAM where a feature could be described by the points with its 2D position (SIFT [4], SURF [5]) or 3D position [6, 7], and also edge segments [8, 9]. But feature extractions from the natural visual scenes were heavily dependent on the environment where the sparse features might be found. These features could be occasionally too few to fully constrain the pose of the robot. Hence the appearance-based SLAM was proposed to represent the recorded images of the environment with prominent features as a whole [10]. Morita et al. reported another novel appearance-based localization approach for outdoor navigation with feature or object learning, recognition, and classification using SVM [11]. However, the usage of rich sensorial information in these appearance-based SLAM solutions has resulted in very time-consuming computation especially for larger-scale environments. To allow real-time operation in more moderately sized environments, one method was proposed to observe the interframe motion of every other corner feature in a visual odometry style [12, 13]. Also, some researchers proposed the method of discovering and incorporating higher level map structure in the forms of lines [14] and planes [15, 16].

Different kinds of maps have been applied in SLAM. Metric maps capture the geometric properties of the environment whereas topological maps describe the connectivity between different locations [17]. Topological maps can represent the environment as a list of the significant places which has simplified the problem of large-scale mapping [18]. However, one limitation of the topological representation was the lack

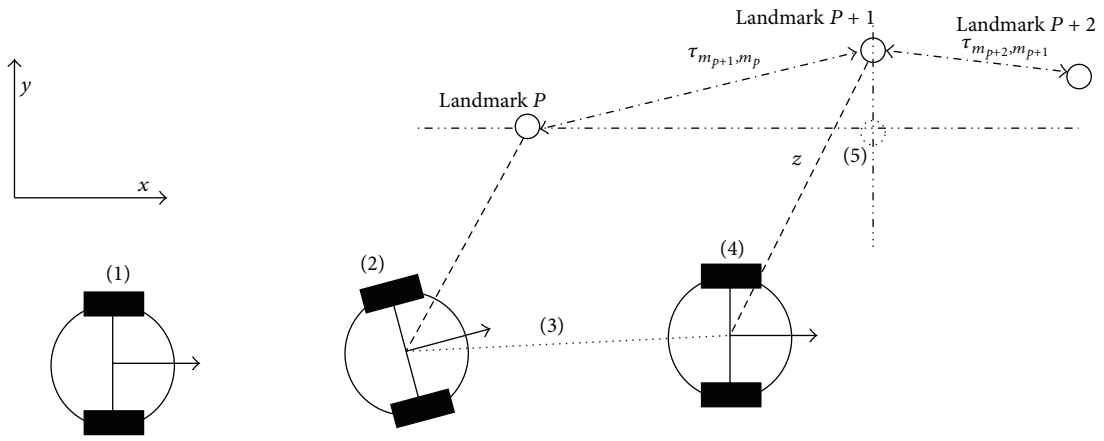


Figure 1: The process of creating the semantic map.

of metric information. So the strategy of mixing the metric and topological information in a single consistent model was proposed [19]. Fernández et al. also developed a hybrid metric-topological algorithm to build a metric map while maintaining a topological graph and to detect loop closures [20]. Thrun and Buecken combined the grid based and topological based methods to map indoor robot environments [21]. Such hybrid algorithms took advantage of the local metric grids for enhanced local planning while avoiding the computation of a complete global grid map. However, these maps are very limited in describing the environment other than distinguishing between occupied and empty areas. In order to explore richer information of the environment, semantic mapping has become a research topic recently. Wolf and Sukhatme proposed a semantic classification method based on HMMs and SVMs to tackle the problem of terrain mapping and activity-based mapping [22]. Ranganathan and Dellaert described a technique to model and recognize the places using objects as the basic semantic concept [23]. Yi et al. proposed a semantic representation and Bayesian model for robot localization using spatial contexts among objects [24, 25]. This paper will take advantage of semantic relationship of features in the visual SLAM framework.

Early work on SLAM was done by Smith et al., where the Extended Kalman Filter (EKF) was applied [26]. Later Doucet et al. introduced the Rao-Blackwellized particle filter (RBPF) as an efficient solution to the SLAM problem which is also called FastSLAM [27]. The Unscented FastSLAM algorithm was then proposed to overcome the drawbacks of FastSLAM where the scaled unscented transformation (SUT) was applied to replace the linearization in the FastSLAM framework [28]. The SLAM solution in this paper will be based on Unscented FastSLAM.

Hence the main contribution of this paper includes a novel Semantic Unscented FastSLAM algorithm to improve accuracy of localization and mapping while maintaining the sparse map for real-time implementation. The semantic relationship and topological metric map are combined to form a new kind of map for SLAM. Few experiments have been carried out for validation of the proposed technique.

The rest of the paper is organized as follows: Section 2 describes the semantic topological metric map and observation model. Framework of the Semantic Unscented FastSLAM is presented in Section 3. The experimental results and discussion are presented in Section 4. The concluding remarks are presented in Section 5.

2. Semantic Topological Metric Map and Observation Model

2.1. Semantic Topological Metric Map. Semantic topological metric map is defined as the combination of the topological metric map and the semantic relationships between the landmarks where the assumption is that such semantic relationships can be represented by some mathematical equations. The spatial semantic relationship between the available landmarks is always invariant with respect to the robot location. Denote the semantic topological metric map as M and the semantic metric relationship as τ. Figure 1 shows the process of creating the semantic topological map. The procedures are summarized as follows. (i) When a robot starts to move and the first landmark is observed, the semantic topological map M only includes the position vector of the m_1, (1). (ii) As the robot moves forward, more landmarks are observed. If there are no semantic relationships between any pair of landmarks, the semantic topological metric map M will be the same as the regular topological metric map. If the number of the observed landmarks is P, the semantic topological map M includes the position vectors of $m_1, \ldots m_p$, (2). (iii) When the robot observes landmark $P+1$, the semantic relationship between landmark $P+1$ and landmark P, τ_{m_{p+1},m_p}, is also found. If all the semantic relationships with the observed landmark $P+1$ are defined as the set $T_{\text{semantic},m_{p+1}}$, the semantic topological metric map M is then updated with the addition of the semantic metric relationship as in (3)-(4):

$$M = \{m_1\}, \tag{1}$$

$$M = \{m_1, \ldots, m_p\}, \tag{2}$$

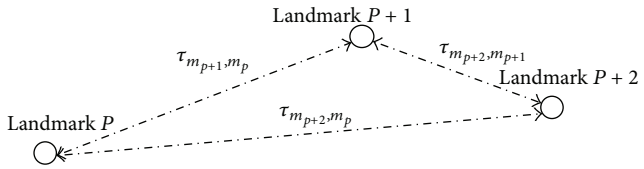


Figure 2: Extension of the semantic topological relationship.

$$M = \left\{m_1, \ldots, m_p, m_{p+1}, T_{\text{semantic},m_{p+1}}\right\}, \quad (3)$$

$$T_{\text{semantic},m_{p+1}} = \left\{\tau_{m_{p+1},m_p}\right\}. \quad (4)$$

(iv) When the robot observes landmark $P+2$, if the semantic relationship between landmarks $P+2$ and $P+1$ is found, then the total semantic topological metric map would be

$$M = \left\{m_1, \ldots, m_p, m_{p+1}, m_{p+2}, T_{\text{semantic},m_{p+2}}\right\}, \quad (5)$$

$$T_{\text{semantic},m_{p+2}} = \left\{\tau_{m_{p+2},m_{p+1}}\right\}. \quad (6)$$

Since more-than-one semantic relationships between different landmarks, (4) and (6), have been observed, the extended new semantic topological relationship will be created, Figure 2, where landmarks $P+2$, $P+1$ and P are associated together. The semantic topological metric map at the time being will become

$$T_{\text{semantic},m_{p+2}} = \left\{\tau_{m_{p+2},m_{p+1}}, \tau_{m_{p+2},m_p}\right\}, \quad (7)$$

where τ_{m_{p+2},m_p} is the extended semantic relationship between landmark $P+2$ and landmark P. When $\tau_{m_{p+2},m_{p+1}}$ has the same semantic relationship as τ_{m_{p+1},m_p} coincidently, we can associate them together as

$$\begin{aligned} &\tau_{m_{p+2},m_{p+1}} = \tau_{m_{p+1},m_p} \\ &\quad \Longrightarrow \tau_{m_{p+2},m_p} = \tau_{m_{p+2},m_{p+1}} = \tau_{m_{p+1},m_p}. \end{aligned} \quad (8)$$

2.2. Semantic Observation Model. A semantic observation model is the observation model of the vision sensor with implicit of the semantic relationships. Hence the semantic observation model consists of not only the metric distance σ and the bearing ω of each landmark, but also the semantic metric relationships between different landmarks. The dimension of the semantic observation model could be $N+1$ where N is the total number of the landmarks observed so far. In this case, the semantic observation model can be represented as

$$\begin{aligned} &z_{\text{semantic}} \\ &= \begin{bmatrix} \sigma \\ \omega \\ \eta_1 \\ \eta_2 \\ \vdots \\ \eta_{N-1} \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{\left(x - m_{N,x}\right)^2 + \left(y - m_{N,y}\right)^2} \\ \tan^{-1}\left(y - \dfrac{m_{N,y}}{x} - m_{N,x}\right) \\ H\left(\tau_{m_N,m_1}, m_{N,x}, m_{N,y}, m_{1,x}, m_{1,y}\right) \\ H\left(\tau_{m_N,m_2}, m_{N,x}, m_{N,y}, m_{2,x}, m_{2,y}\right) \\ \vdots \\ H\left(\tau_{m_N,m_{N-1}}, m_{N,x}, m_{N,y}, m_{N-1,x}, m_{N-1,y}\right) \end{bmatrix} \\ &= \varphi\left(X, m, T_{\text{sementic},m_N}\right), \end{aligned} \quad (9)$$

where $H(*)$ is the mathematical expression of the semantic metric relationship, (x, y) is the coordinates of the current robot pose, and $(m_{N,x}, m_{N,y})$ is the coordinates of the landmark N observed at the current time period. $\tau_{m_N,m_1}, \tau_{m_N,m_2}, \ldots, \tau_{m_N,m_{N-1}}$ are the series of the possible semantic metric relationships associated with m_N. The position vector of the robot is defined as $X = [x, y, \theta]^T$. m is defined as the position set of all the landmarks observed at the current time as follows:

$$m = \left(m_1, \ldots, m_N\right)^T. \quad (10)$$

3. Semantic Unscented FastSLAM

Semantic unscented FastSLAM partitions the SLAM posterior into a localization problem and independent landmark position estimation problem conditioned on the robot pose estimate and the semantic metric relationships between the landmarks as follows:

$$\begin{aligned} p\left(X_t, m_t, r_t \mid z_{t,\text{semantic}}, u_t\right) &= p\left(X_t \mid z_{t,\text{semantic}}, u_t\right) \\ &\cdot \prod_{j=1}^{N} p\left(m_{j,t}, T_{\text{semantic},m_j} \mid X_t, z_{t,\text{semantic}}\right), \end{aligned} \quad (11)$$

where X_t is the robot pose at time t and r denotes the full semantic metric map at the current time period as follows:

$$r_t = \left(T_{\text{semantic},m_1}, T_{\text{semantic},m_2}, \ldots, T_{\text{semantic},m_N}\right)^T. \quad (12)$$

Suppose the control vector of the robot is $u = [v, w]^T$ where v and w represent the linear and angular velocities of the

robot. According to the kinematics of the wheeled mobile robot [29], the motion model is represented as follows:

$$X_t = X_{t-1} + \begin{bmatrix} -\frac{v_t}{w_t}\sin\theta_{t-1} + \frac{v_t}{w_t}\sin\left(\theta_{t-1} + w_t\Delta t\right) \\ \frac{v_t}{w_t}\cos\theta_{t-1} - \frac{v_t}{w_t}\cos\left(\theta_{t-1} + w_t\Delta t\right) \\ w_t\Delta t \end{bmatrix} = f\left(u_t, X_{t-1}\right). \tag{13}$$

3.1. Robot Pose Estimation. Since particle filter is incorporated into the FastSLAM frame, the following derivation will be associated with only one particle as an example. Then the robot pose at time t for a kth particle can be estimated as

$$p\left(X_t^{[k]} \mid z_{t,\text{semantic}}, u_t\right) \sim p\left(X_t^{[k]} \mid X_{t-1}^{[k]}, u_t\right) p\left(X_{t-1}^{[k]} \mid z_{t-1,\text{semantic}}, u_{t-1}\right), \tag{14}$$

where $p(X_{t-1}^{[k]} \mid z_{t-1,\text{semantic}}, u_{t-1})$ is represented by a Gaussian with the mean $X_{t-1}^{[k]}$ and covariance $P_{t-1}^{[k]}$. The $p(X_t^{[k]} \mid X_{t-1}^{[k]}, u_t)$ can be predicted in the following according to the motion model of the robot. In order to integrate the robot pose and the map update, the state vector is augmented with a control input and the observation vector as

$$X_{t-1}^{a[k]} = \left[X_{t-1}^{[k]}, 0, 0\right]^T, \quad P_{t-1}^{a[k]} = \begin{bmatrix} P_{t-1}^{[k]} & 0 & 0 \\ 0 & Q_t & 0 \\ 0 & 0 & R_t \end{bmatrix}, \tag{15}$$

where $X_{t-1}^{a[k]}$ is the augmented state vector, Q_t is the motion noise covariance and R_t is the observation noise covariance, and $P_{t-1}^{a[k]}$ is the augmented covariance matrix.

In order to apply the unscented transformation, a symmetric set of $2n + 1$ sigma points (n is the dimension of the augmented state vector) need to be extracted first as follows [30]:

$$\begin{aligned} \chi_{t-1}^{a[0][k]} &= X_{t-1}^{a[k]}, \\ \chi_{t-1}^{a[i][k]} &= X_{t-1}^{a[k]} + \left(\sqrt{(n+\lambda)\,P_{t-1}^{[k]}}\right)_i, \quad \text{for } i = 1, \ldots, n, \\ \chi_{t-1}^{a[i][k]} &= X_{t-1}^{a[k]} - \left(\sqrt{(n+\lambda)\,P_{t-1}^{[k]}}\right)_{i-n}, \\ &\qquad \text{for } i = n+1, \ldots, 2n, \end{aligned} \tag{16}$$

where the subscript i means the ith column of a matrix. The λ is computed by $\lambda = \alpha^2(n+\kappa) - n$ and α is a small number to avoid the sampling nonlocal effects for high nonlinearities. κ is a scaling parameter determining how far the sigma points are separated from the mean value. Each sigma point $\chi_{t-1}^{a[i][k]}$ contains the robot pose, control noise, and semantic observation noise components as

$$\chi_{t-1}^{a[i][k]} = \begin{bmatrix} \chi_{t-1}^{[i][k]} \\ \chi_t^{u[i][k]} \\ \chi_t^{z[i][k]} \end{bmatrix}. \tag{17}$$

So the prediction of the robot pose can be derived by passing the above sigma points through the motion model, f in (13). The transformed sigma points of the robot pose, $\overline{\chi}_t^{[i][k]}$, are calculated as

$$\overline{\chi}_t^{[i][k]} = f\left(u_t^{[k]} + \chi_t^{u[i][k]}, \chi_{t-1}^{[i][k]}\right), \tag{18}$$

where the current control vector is the sum of the $u_t^{[k]}$ and the control noise component $\chi_t^{u[i][k]}$ of each sigma point. Then the prediction of the robot pose can be calculated as

$$\begin{aligned} X_{t|t-1}^{[k]} &= \sum_{i=0}^{2n} \omega_b^{[i]} \overline{\chi}_t^{[i][k]}, \\ P_{t|t-1}^{k} &= \sum_{i=1}^{2n} \omega_c^{[i]} \left(\overline{\chi}_t^{[i][k]} - X_{t|t-1}^{[k]}\right)\left(\overline{\chi}_t^{[i][k]} - X_{t|t-1}^{[k]}\right)^T. \end{aligned} \tag{19}$$

The weights are calculated by the following equations:

$$\begin{aligned} \omega_b^{[0]} &= \frac{\lambda}{n+\lambda}, \\ \omega_c^{[0]} &= \frac{\lambda}{n+\lambda} + \left(1 - \alpha^2 + \beta\right), \\ \omega_c^{[i]} &= \omega_b^{[i]} = \frac{1}{2\,(n+\lambda)}, \quad \text{for } i = 1, \ldots, 2n, \end{aligned} \tag{20}$$

where the weight $\omega_c^{[i]}$ is used to compute the mean of the predicted robot pose, and the weight $\omega_b^{[i]}$ is used to recover the covariance of the Gaussian. The parameter β is used to incorporate the knowledge of the higher order moments of the posterior distribution.

Suppose the jth landmark and its semantic relationships are observed; the transformed sigma points of the semantic observation vector can be derived as

$$\overline{N}_t^{[i][k]} = \varphi\left(\overline{\chi}_t^{[i][k]}, m_{t-1}^{[k]}, T_{\text{semantic},m_j}^{[k]}\right) + \chi_t^{z[i][k]}, \tag{21}$$

where the semantic metric relationships, $T_{\text{semantic},m_j}^{[k]}$, are included in the semantic observation model $\varphi(\cdot)$ in (9) for robot pose update. So this new update will result in the improvement of robot localization. Then the prediction of the semantic observation vector can be calculated as

$$\widehat{n}_t^{[k]} = \sum_{i=0}^{2n} \omega_b^{[i]} \overline{N}_t^{[i][k]}. \tag{22}$$

The Kalman gain can then be obtained by the following equations as usual:

$$S_t^{[k]} = \sum_{i=0}^{2n} \omega_c^{[i]} \left(\overline{N}_t^{[i][k]} - \hat{n}_t^{[k]}\right)\left(\overline{N}_t^{[i][k]} - \hat{n}_t^{[k]}\right)^T,$$
$$\eta_t^{x,n[k]} = \sum_{i=0}^{2n} \omega_c^{[i]} \left(\overline{\chi}_t^{[i][k]} - X_{t|t-1}^{[k]}\right)\left(\overline{N}_t^{[i][k]} - \hat{n}_t^{[k]}\right)^T, \quad (23)$$
$$K_t^{[k]} = \eta_t^{x,n[k]} \left(S_t^{[k]}\right)^{-1},$$

where $S_t^{[k]}$ is the innovation covariance and $\eta_t^{x,n[k]}$ is the cross-covariance.

Therefore, the mean and covariance of the robot pose are estimated at the time period t by

$$X_t^{[k]} = X_{t|t-1}^{[k]} + K_t^{[k]} \left(z_t - \hat{n}_t^{[k]}\right), \quad (24)$$
$$P_t^{[k]} = P_{t|t-1}^{[k]} - K_t^{[k]} S_t^{[k]} \left(K_t^{[k]}\right)^T. \quad (25)$$

3.2. Landmark Position Estimate with Semantic Constraints. For the observed landmark j, the probability of the landmark position estimate can be represented as

$$p\left(m_{j,t}^{[k]}, T_{\text{semantpic},m_j} \mid X_t^{[k]}, z_{t,\text{semantic}}^{[k]}\right) \sim p\left(z_{t,\text{semantic}}^{[k]} \mid X_t^{[k]}, m_{j,t}^{[k]}, T_{\text{semantpic},m_j}\right) \cdot p\left(m_{j,t-1}^{[k]}, T_{\text{semantpic},m_j} \mid X_{t-1}^{[k]}, z_{t-1,\text{semantic}}^{[k]}\right), \quad (26)$$

where the probability $p(m_{j,t-1}^{[k]}, T_{\text{semantpic},m_j} \mid X_{t-1}^{[k]}, z_{t-1,\text{semantic}}^{[k]})$ is represented by a Gaussian with the mean $m_{j,t-1}^{[k]}$ and covariance $\Sigma_{j,t-1}^{[k]}$. $p(z_{t,\text{semantic}}^{[k]} \mid X_t^{[k]}, m_{j,t}^{[k]}, T_{\text{semantpic},m_j})$ will be derived in the following. Likewise, the sigma points of the observed landmark position, m_j, are initialized as

$$\chi_{m,t-1}^{[0][k]} = m_{j,t-1}^{[k]},$$
$$\chi_{m,t-1}^{[i][k]} = m_{j,t-1}^{[k]} + \left(\sqrt{(n+\lambda)\,\Sigma_{j,t-1}^{[k]}}\right)_i, \quad \text{for } i = 1, \ldots, n, \quad (27)$$
$$\chi_{m,t-1}^{[i][k]} = m_{j,t-1}^{[k]} - \left(\sqrt{(n+\lambda)\,\Sigma_{j,t-1}^{[k]}}\right)_{i-n}, \quad \text{for } i = n+1, \ldots, 2n.$$

The transformed sigma points of the landmark position estimation with semantic relationships can be derived as

$$\overline{Z}_{t,\text{semantic}}^{[i][k]} = \varphi\left(X_t^{[k]}, \chi_{m,t-1}^{[i][k]}, T_{\text{semantic},m_j}^{[k]}\right), \quad (i = 0, \ldots, 2n), \quad (28)$$

where $\varphi(\cdot)$ is the observation model in (9), $X_t^{[k]}$ is the current estimation of the robot pose in (24). Hence the predicted semantic observation vector, $\hat{z}_{t,\text{semantic}}^{[k]}$, is

$$\hat{z}_{t,\text{semantic}}^{[k]} = \sum_{i=0}^{2n} \omega_b^{[i]} \overline{Z}_{t,\text{semantic}}^{[i][k]}. \quad (29)$$

Then the Kalman gain $\overline{K}_{t,\text{semantic}}^{[k]}$ is calculated as follows:

$$\overline{S}_{t,\text{semantic}}^{[k]} = \sum_{i=0}^{2n} \omega_c^{[i]} \left(\overline{Z}_{t,\text{semantic}}^{[i][k]} - \hat{z}_{t,\text{semantic}}^{[k]}\right) \cdot \left(\overline{Z}_{t,\text{semantic}}^{[i][k]} - \hat{z}_{t,\text{semantic}}^{[k]}\right)^T + R_t,$$
$$\overline{\Sigma}_{t,\text{semantic}}^{[k]} = \sum_{i=0}^{2n} \omega_c^{[i]} \left(\chi_m^{[i][k]} - m_{j,t-1}^{[k]}\right) \cdot \left(\overline{Z}_{t,\text{semantic}}^{[i][k]} - \hat{z}_{t,\text{semantic}}^{[k]}\right)^T, \quad (30)$$
$$\overline{K}_{t,\text{semantic}}^{[k]} = \overline{\Sigma}_{t,\text{semantic}}^{[k]} \left(\overline{S}_{t,\text{semantic}}^{[k]}\right)^{-1}.$$

Note that the weights $\omega_b^{[i]}$ and $\omega_c^{[i]}$ are the same as (20). Finally, the mean $m_{j,t}^{[k]}$ and the covariance $\Sigma_{j,t}^{[k]}$ of the jth landmark position are updated by

$$m_{j,t}^{[k]} = m_{j,t-1}^{[k]} + \overline{K}_{t,\text{semantic}}^{[k]} \left(z_{t,\text{semantic}} - \hat{z}_{t,\text{semantic}}^{[k]}\right),$$
$$\Sigma_{j,t}^{[k]} = \Sigma_{j,t-1}^{[k]} - \overline{K}_{t,\text{semantic}}^{[k]} \overline{S}_{t,\text{semantic}}^{[k]} \left(\overline{K}_{t,\text{semantic}}^{[k]}\right)^T. \quad (31)$$

Note that $z_{t,\text{semantic}}$ includes the true observation of the relative position of the landmark and the robot and the associated semantic relationships with this landmark. These observation values are obtained from the image process of the vision sensor data. If more landmarks are observed at one time, the derivation would be similar except that more semantic relationships would be included in the observation model.

As mentioned at the beginning, all the above derivation is with respect to the particle k. Then the traditional resampling procedure will be taken, and the robot pose and the landmark positions will be estimated finally.

4. Experiments and Discussions

4.1. Experiment Procedures. The platform used in the experiments was a Pioneer 3-DX robot equipped with a binocular camera system. The camera was the only exteroceptive sensor to recover the representation of the environment. The sampling period was 0.5 seconds. The proposed technique has been evaluated by three different types of the experiments. In Experiment 1, the robot moved along a simple rectangular trajectory (8 m × 14 m) in a neat lab environment. The environment in Experiment 2 was a regular office area that was more general to most indoor service robots to verify the superior performance of the Semantic Unscented FastSLAM.

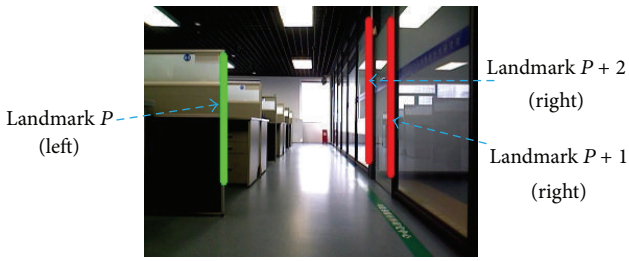


Figure 3: The observation result of vision system.

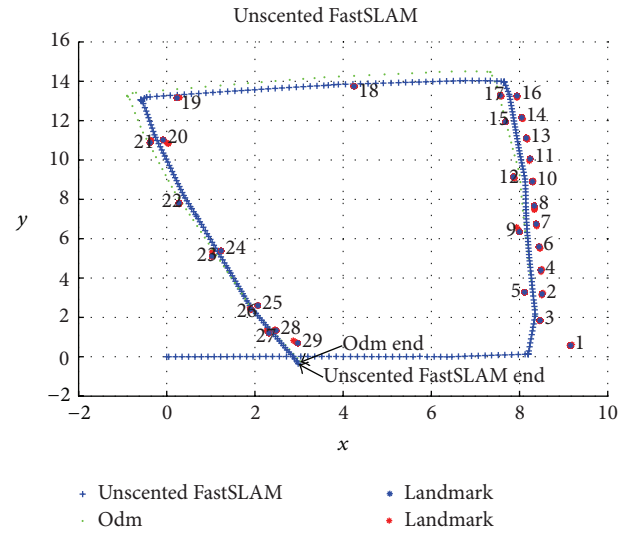
Experiment 3 was conducted in a messy environment where the robot had to move along a zig-zag path to go through aisles.

In the experiments, three kinds of the semantic metric relationships were found. One semantic relationship was that the new observed landmark and another landmark existing in the previous map were both along the x-axis (x-line). The second semantic relationship was that two landmarks were both along y-axis (y-line). Such two kinds of semantic relationships are denoted by $\{x\text{-line}, y\text{-line}\}$. The third one was that three landmarks were collinear such as the walls of neighboring cubes in an office. Suppose m_p, m_s, and m_q are three landmarks with the above semantic relationships; the semantic observation model can be represented, respectively, as

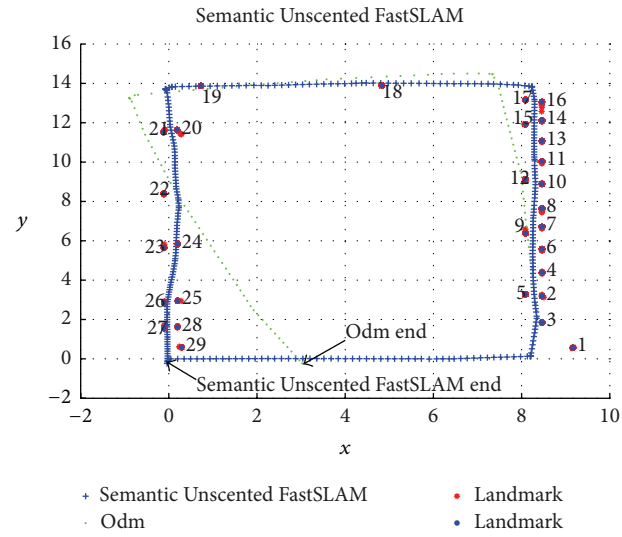
$$\begin{aligned} H\left(\tau_{m_p,m_s}, m_{p,x}, m_{p,y}, m_{s,x}, m_{s,y}\right)\Big|_{(\tau_{m_p,m_s}=x\text{-line})} &= m_{p,y} \\ &- m_{s,y}, \\ H\left(\tau_{m_p,m_s}, m_{p,x}, m_{p,y}, m_{s,x}, m_{s,y}\right)\Big|_{(\tau_{m_p,m_s}=y\text{-line})} &= m_{p,x} \\ &- m_{s,x}, \\ H\left(\tau_{m_p,m_s,m_q}, m_{p,x}, m_{p,y}, m_{s,x}, m_{s,y}, m_{q,x},\right. & \\ \left. m_{q,y}\right)\Big|_{(\tau_{m_p,m_s,m_q}=\text{collinear})} &= \left(m_{p,y} - m_{s,y}\right)\left(m_{p,x}\right. \\ \left. - m_{q,x}\right) - \left(m_{p,y} - m_{q,y}\right)\left(m_{p,x} - m_{s,x}\right). & \end{aligned} \tag{32}$$

4.2. Experiment Results and Discussions

Experiment 1. Figure 3 shows one image taken by the vision sensor on the robot with three landmarks P, $P+1$, and $P+2$ where the landmarks $P+2$ and $P+1$ were located along the x-axis. This semantic relationship will be applied for localization and mapping. Figure 4 shows the comparison of the system performance using the Unscented FastSLAM (Figure 4(a)) and the proposed method (Figure 4(b)). As shown in Figure 4(a), the error of robot pose became larger especially after the robot was turning. This error could not be corrected by the loop closure detection because all the landmarks observed after turning have not been observed before. In Figure 4(b), the localization error has been eliminated greatly after the semantic topological metric map was



(a) Experiment result using Unscented FastSLAM

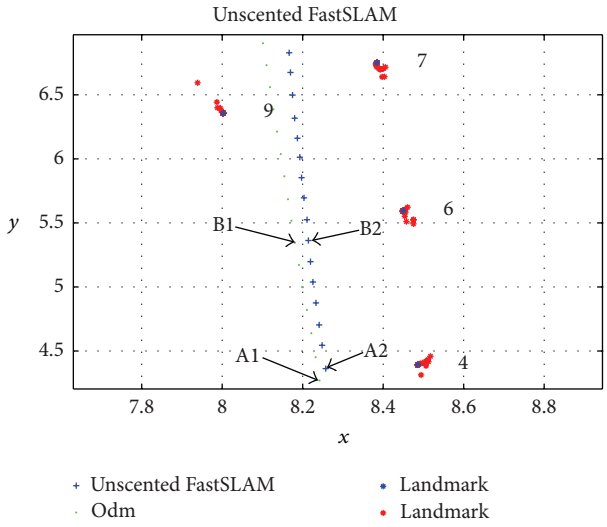


(b) Experiment result using Semantic Unscented FastSLAM

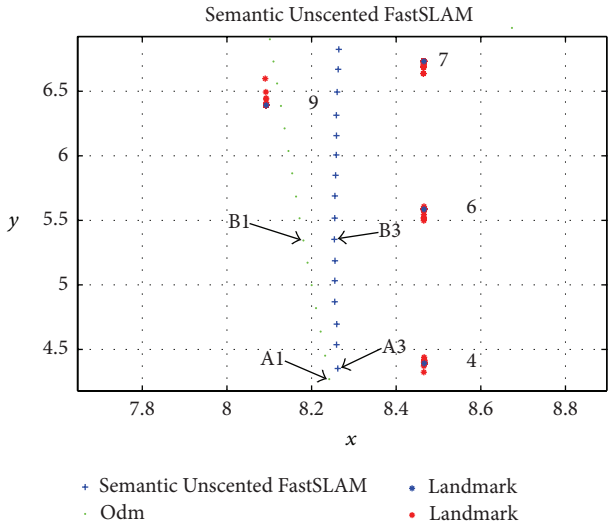
Figure 4: Comparison of SLAM results, Experiment 1.

applied. Figure 5 is the partially enlarged view of Figure 4 where A1 and B1 are the estimations by odometer only, A2 and B2 are the estimation from Unscented FastSLAM, and A3 and B3 are the estimation from the proposed Semantic Unscented FastSLAM. When Landmark #6 was observed by the robot at the first time, it was also found that Landmark #6 has the semantic topological relationship "y-line" with the landmarks #4, #2, and #3 in the existing map. Hence this semantic relationship has resulted in much better robot pose estimation, B3 in Figure 5(b), which has pulled the dead reckoning estimate B1 back from the deviation comparing with B2 without taking advantage of semantic relationships.

Experiment 2. Figure 6 shows the experimental environment in Experiment 2 where the reference trajectory started from the circle and ended at the same point after a complex



(a) Experiment result using Unscented FastSLAM



(b) Experiment result using Semantic Unscented FastSLAM

FIGURE 5: Partially enlarged view of SLAM results in Experiment 1.

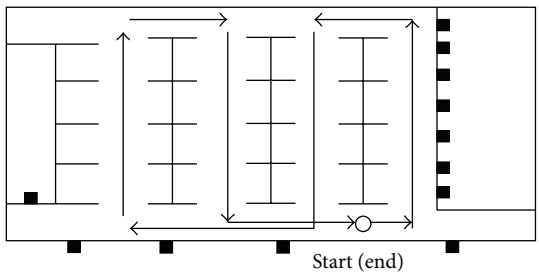
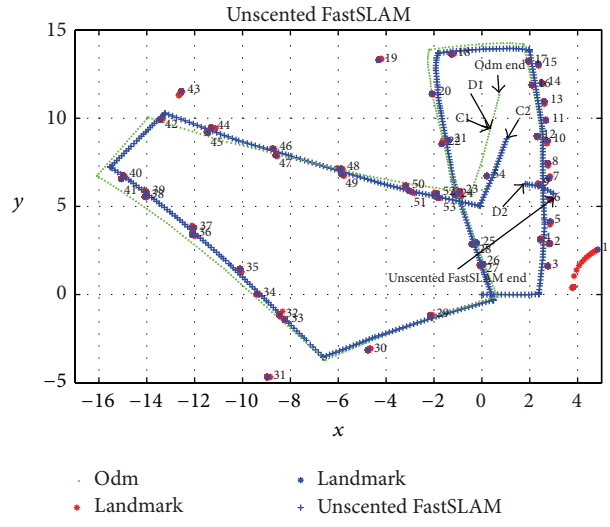
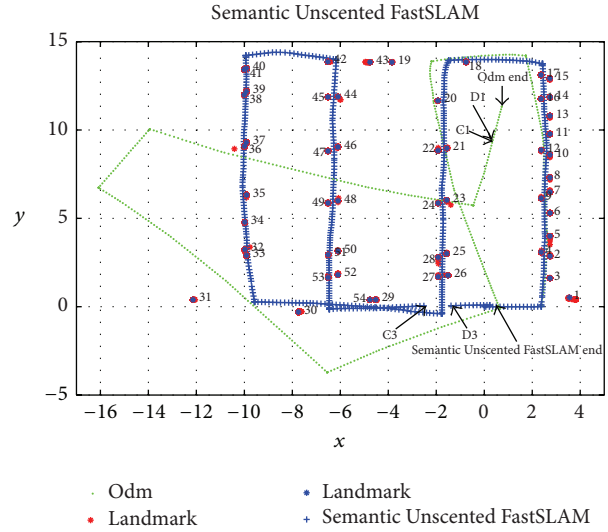


FIGURE 6: Floor map of the office environment, Experiment 2.

surveillance along the arrow directions. The start point was defined as the origin of the inertial frame $(0, 0)$. It is worth noting that this office was composed of a few cubes that were higher than the robot. Hence when the robot moved along the reference trajectory, most landmarks could not be observed more than once before the robot was close to the end point.



(a) Experiment result using Unscented FastSLAM



(b) Experiment result using Semantic Unscented FastSLAM

FIGURE 7: Comparison of SLAM results, Experiment 2.

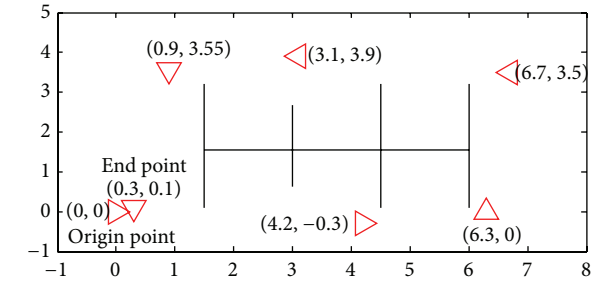


FIGURE 8: Map of the office environment with reference path, Experiment 3.

The experimental results using the Unscented FastSLAM and the proposed Semantic Unscented FastSLAM are shown in Figure 7. In this experiment, when the robot moved close to the end point, Landmark #1 should be observed after a long

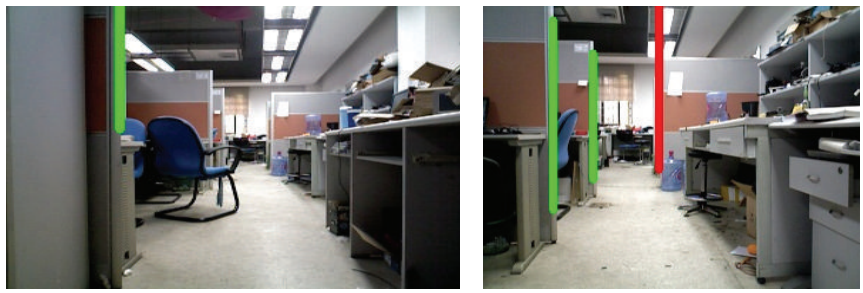
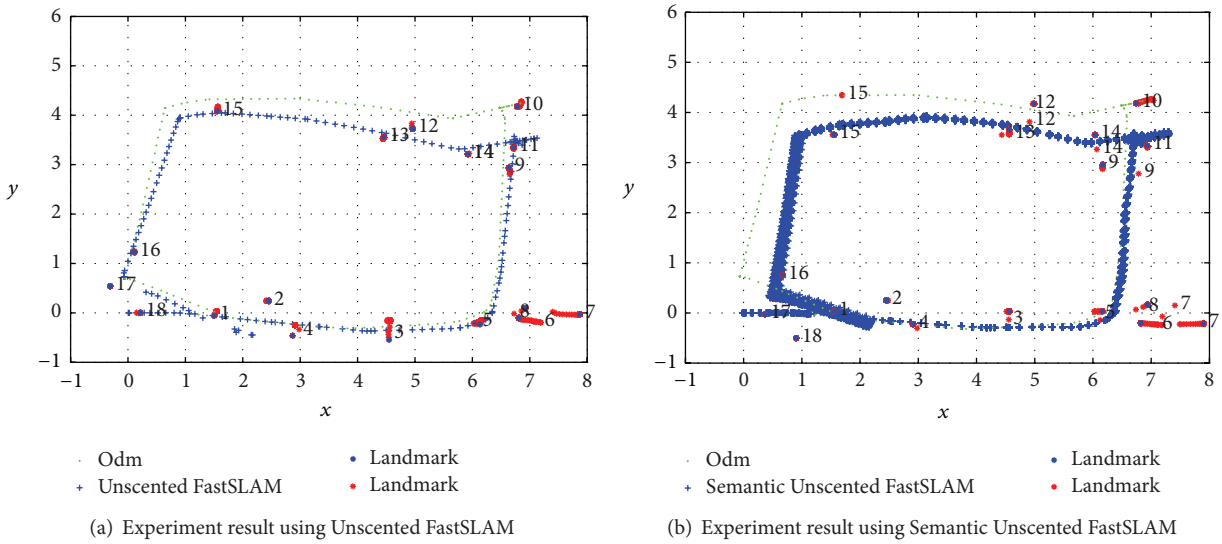


Figure 9: Two pictures with three collinear landmarks, Experiment 3.



(a) Experiment result using Unscented FastSLAM

(b) Experiment result using Semantic Unscented FastSLAM

Figure 10: Comparison of SLAM results, Experiment 3.

period for the loop closure detection. The estimation of the end point, C1, by odometer only was far away from the real end point. As shown in Figure 7(a), the end point estimated by the Unscented FastSLAM before the loop closure detection of Landmark #1, C2, was better but still had a huge error. This error was too large to be corrected by the loop closure detection (see D2 for the estimation after loop closure detection). Figure 7(b) shows that the end point estimated by the proposed Semantic Unscented FastSLAM before the loop closure detection, C3, was much smaller because of the semantic updates in the algorithm. Therefore, after the loop closure detection, the error was reduced close to the reference point (see D3 for the estimation by Semantic Unscented FastSLAM in Figure 7(b)).

Experiment 3. The experiment environment is shown in Figure 8 where small triangles represent a few locations along the reference path, and the solid line represents the wall of cubes. Notice that the reference path is not straightforward during each aisle because the aisle had irregular width and the robot also needs to avoid chairs and boxes on both sides of the aisle. Figure 9 shows an example of two pictures captured by the camera where three green landmarks were detected as collinear relationship. Figure 10 illustrates the performance of the surveillance robot in Experiment 3. As shown in Figure 10(b), the locations of robot and landmarks are much closer to the reference path using the proposed Semantic Unscented FastSLAM than without considering semantic relationships (Figure 10(a)).

5. Conclusions

This paper has proposed a vision-based Semantic Unscented FastSLAM for mobile robot. The semantic relationship is combined with the traditional topological metric map to improve the accuracy of localization and mapping. Experiments were conducted to verify that the Semantic Unscented FastSLAM is more robust and applicable to more general indoor autonomous surveillance.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research is currently supported by the National Science and Technology Ministry of China under Grant no.

2013BAK01B02 and was supported by the State Key Laboratory of Robotics and System (HIT) under SKLRS-2011-ZD-04. The partial support of the National Natural Science Foundation of China (no. 61273335) and National High-Tech Research and Development Program of China (863 Program, no. 2015AA042303) is also appreciated.

References

[1] R. Sim, P. Elinas, M. Griffin, A. Shyr, and J. J. Little, "Design and analysis of a framework for real-time vision-based SLAM using Rao-Blackwellised particle filters," in *Proceedings of the 3rd Canadian Conference on Computer and Robot Vision (CRV '06)*, p. 21, IEEE, June 2006.

[2] R. Sim and J. J. Little, "Autonomous vision-based exploration and mapping using hybrid maps and Rao-Blackwellised particle filters," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '06)*, pp. 2082–2089, October 2006.

[3] D. Schleicher, L. M. Bergasa, R. Barea, E. López, and M. Ocaña, "Real-time simultaneous localization and mapping using a wide-angle stereo camera and adaptive patches," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '06)*, pp. 2090–2095, October 2006.

[4] S. Frintrop and P. Jensfelt, "Attentional landmarks and active gaze control for visual SLAM," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1054–1065, 2008.

[5] M. Magnabosco and T. P. Breckon, "Cross-spectral visual simultaneous localization and mapping (SLAM) with sensor handover," *Robotics and Autonomous Systems*, vol. 61, no. 2, pp. 195–208, 2013.

[6] B. Williams, P. Smith, and I. Reid, "Automatic relocalisation for a single-camera simultaneous localisation and mapping system," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '07)*, pp. 2784–2790, April 2007.

[7] D. Chekhlov, M. Pupilli, W. Mayol, and A. Calway, "Robust real-time visual SLAM using scale prediction and exemplar based feature description," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–7, Minneapolis, Minn, USA, June 2007.

[8] E. Eade and T. Drummond, "Edge landmarks in monocular SLAM," *Image and Vision Computing*, vol. 27, no. 5, pp. 588–596, 2009.

[9] C. Zhou, Y. Wei, and T. Tan, "Mobile robot self-localization based on global visual appearance features," in *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1271–1276, September 2003.

[10] B. Bacca, J. Salvi, and X. Cufí, "Appearance-based SLAM for mobile robots," in *Proceedings of the 12th International Conference of the Catalan Association for Artificial Intelligence (CCIA '09)*, pp. 55–64, Cardona, Spain, October 2009.

[11] H. Morita, M. Hild, J. Miura, and Y. Shirai, "Panoramic view-based navigation in outdoor environments based on support vector learning," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '06)*, pp. 2302–2307, October 2006.

[12] B. Williams and I. Reid, "On combining visual SLAM and visual odometry," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '10)*, pp. 3494–3500, May 2010.

[13] J. Martínez-Carranza and A. Calway, "Efficient visual odometry using a structure-driven temporal map," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '12)*, pp. 5210–5215, IEEE, Saint Paul, Minn, USA, May 2012.

[14] A. P. Gee, D. Chekhlov, A. Calway, and W. Mayol-Cuevas, "Discovering higher level structure in visual SLAM," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 980–990, 2008.

[15] E. Fernandez-Moral, W. Mayol-Cuevas, V. Arevalo, and J. Gonzalez-Jimenez, "Fast place recognition with plane-based maps," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '13)*, pp. 2719–2724, May 2013.

[16] J. Kwon and K. M. Lee, "Monocular SLAM with locally planar landmarks via geometric rao-blackwellized particle filtering on lie groups," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 1522–1529, June 2010.

[17] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[18] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Visual topological SLAM and global localization," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '09)*, pp. 4300–4305, Kobe, Japan, May 2009.

[19] K. Konolige, E. Marder-Eppstein, and B. Marthi, "Navigation in hybrid metric-topological maps," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '11)*, pp. 3041–3047, May 2011.

[20] L. Fernández, L. Payá, O. Reinoso, A. Gil, and D. Valiente, "Visual hybrid SLAM: an appearance-based approach to loop closure," in *ROBOT2013: First Iberian Robotics Conference*, vol. 252 of *Advances in Intelligent Systems and Computing*, pp. 693–701, Springer, 2014.

[21] S. Thrun and A. Buecken, "Integrating grid-based and topological maps for mobile robot navigation," in *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI '96)*, pp. 944–950, August 1996.

[22] D. F. Wolf and G. S. Sukhatme, "Semantic mapping using mobile robots," *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 245–258, 2008.

[23] A. Ranganathan and F. Dellaert, "Semantic modeling of places using objects," in *Proceedings of the Robotics: Science and Systems Conference*, pp. 27–30, Atlanta, Ga, USA, June 2007.

[24] C. Yi, I. H. Suh, G. H. Lim, and B.-U. Choi, "Active-semantic localization with a single consumer-grade camera," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '09)*, pp. 2161–2166, October 2009.

[25] C. Yi, I. H. Suh, G. H. Lim, and B.-U. Choi, "Bayesian robot localization using spatial object contexts," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '09)*, pp. 3467–3473, IEEE, St. Louis, Mo, USA, October 2009.

[26] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *Autonomous Robot Vehicles*, pp. 167–193, Springer, Berlin, Germany, 1990.

[27] A. Doucet, N. De Freitas, K. Murphy, and S. Russell, "Rao-blackwellised particle filtering for dynamic Bayesian networks," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI '00)*, pp. 176–183, 2000.

[28] C. Kim, R. Sakthivel, and W. K. Chung, "Unscented FastSLAM: a robust and efficient solution to the SLAM problem," *IEEE Transactions on Robotics*, vol. 24, no. 4, pp. 808–820, 2008.

[29] G. Campion, G. Bastin, and B. Dandrea-Novel, "Structural properties and classification of kinematic and dynamic models of wheeled mobile robots," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 1, pp. 47–62, 1996.

[30] S. Julier, J. Uhlmann, and H. F. Durrant-Whyte, "A new method for the nonlinear transformation of means and covariances in filters and estimators," *IEEE Transactions on Automatic Control*, vol. 45, no. 3, pp. 477–482, 2000.