Vision-Based Semantic-Map Building and Localization

Seungdo Jeong[1], Jounghoon Lim[2], Il Hong Suh[2], and Byung-Uk Choi[2]

[1] Department of Electrical and Computer Engineering, Hanyang University
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791 Korea
sdjeong@mlab.hanyang.ac.kr
[2] School of Information and Communications, Hanyang University,
lightinday@hanafos.com, {ihsuh, buchoi}@hanyang.ac.kr

Abstract. A semantic-map building method is proposed to localize a robot in the semantic-map. Our semantic-map is organized by using SIFT feature-based object representation. In addition to semantic map, a vision-based relative localization is employed as a process model of extended Kalman filters, where optical flows and Levenberg-Marquardt least square minimization are incorporated to predict relative robot locations. Thus, robust SLAM performances can be obtained even under poor conditions in which localization cannot be achieved by classical odometry-based SLAM.

1 Introduction

To navigate and plan a path in an environment, the intelligent mobile robot has to be able to build a map of the environment and to recognize its location autonomously. SLAM(Simultaneous Localization And Mapping) is a popular technique to accurately localize the robot and simultaneously build a map of the environment. Numerous studies of SLAM have been performed, as this has been one of the important issues in the intelligent mobile robot community for a long time. Thus, a number of sensors and algorithms have been proposed to realize the SLAM technique. Since the 1990s, methods using extended Kalman filters have been focused on by a number of researchers interested in SLAM [3, 8]. Most algorithms using extended Kalman filters combine two methods. One is relative localization, which is the method used to compute the current position with respect to an initial location. The other is the method using a map composed of landmarks.

Odometry is often used for relative localization. However, it has many errors caused not only by systematic factors such as a difference in wheel diameter, inaccurate gauging of wheel size, and others, but also non-systematic factors such as slip, poor condition of the floor, etc. [4] proposed an algorithm to reduce such systematic errors. Their algorithm consists of three steps: odometry error modeling, error parameter estimation using the PC-method, and estimation of the covariance matrix. Even so, it is hard to overcome the error of estimation caused by non-systematic factors.

B. Gabrys, R.J. Howlett, and L.C. Jain (Eds.): KES 2006, Part I, LNAI 4251, pp. 559–568, 2006.

The Harris three-dimensional vision system DROID is another algorithm for relative localization [5]. This system uses the visual motion of image corner features for 3D reconstruction. Kalman filters are used for tracking features, and from locations of the tracked image features, DROID determines both the camera motion and the 3D position of the features. Ego-motion determination by matching image features is generally very accurate in the short to medium term.

Building the map of the environment is another issue with SLAM. The map consists of landmarks which are used for robot localization. Davision uses three-dimensional corner points as landmarks, which are obtained by a Harris corner detector and stereo matching [5]. Se uses keypoints obtained from SIFT(Scale Invariant Feature Transform) [6, 9]. However, simple coordinates or features within the image are not enough to facilitate interaction with humans.

In this work, we use known objects as landmarks through object recognition and then build the semantic-map with these recognized object landmarks. This semantic-map is very helpful for interaction between a human and the robot. We also use vision-based relative localization process as the process model of our extended Kalman filter. This approach is robust enough for environments which cannot be supported by encoders that measure values such as the number of rotations of the robot's wheels.

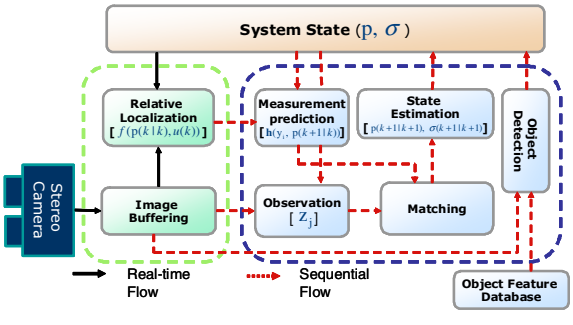

Fig. 1. Block diagram of the proposed method

2 OFM-Based Robot Localization and Semantic-Map Building

In this work, we propose a three-dimensional object feature model (OFM) with essential properties and propose a method to use the OFM for improving the performance of vision-based SLAM. Figure 1 shows the block diagram of our SLAM method using a three-dimensional OFM.

We use images taken by the stereo camera as the only sensor information in this work. Our system is composed of three parts: the real-time relative localization part which estimates robot location in real-time, the landmark recognition

part which builds up and/or observes landmarks by using SIFT-based object recognition, and the data fusion part which combines two observation data using an extended Kalman filter.

2.1 Real-Time Relative Localization

Image-based relative localization is a method that applies the motion between corresponding points within consecutive image sequences to localize the robot. Thus, matching between feature points is a crucial factor for localization performance.

Tracking of Corner Features. To track the camera motion, it is necessary to obtain the feature points in the current frame corresponding to the feature points in the previous frame. Note that the corner points satisfy the local smoothness constraint. Lucas-Kanade optical flow performs well in tracking corner points with that property [7]. However, Lucas-Kanade optical flow is best suited to small motion tracking. It is not sufficient to follow the movement of the robot. Thus, pyramid Lucas-Kanade (PLK) optical flow using a Gaussian image pyramid is used for relative localization, where it is known to track a relatively wide area.

Localization. The feature points extracted from the initial frame are used as 3D point landmarks. Note that we use the stereo camera as the input sensor, and we obtain a stereo image pair. The 3D coordinates of the extracted feature point are calculated using the disparity of the stereo camera images. The stereo camera is previously calibrated. When we know the exact relationship between the 3D coordinates of the landmark and the corresponding 2D coordinates which are projected to the image, the projection matrix indicates the mapping relationship between points of 3D space and points of the image.

The projection matrix is composed of two parts. One is the intrinsic parameter matrix which includes internal information of the camera. This matrix represents the relationship between the camera coordinate system and the image coordinate system. The other part of the projection matrix is the extrinsic parameter matrix. This matrix represents the translational and rotational relationship between the 3D space coordinate system and the camera coordinate system.

Therefore, if the projection matrix is estimated with coordinates of corresponding points in 3D space and in the 2D image, we can obtain the motion of the camera and the relative location of the robot. It is a non-linear problem to estimate the projection matrix with numerous corresponding points. In this work, we estimate the rotation and translation components using the Levenberg-Marquardt least square minimization (LM LSM) method.

2.2 Extended Kalman Filter-Based SLAM Using Object Landmarks

In most previous research works on SLAM, landmarks were composed of lines of the environment. They were obtained using a range finder or feature points

of camera images. These maps only involve coordinates of feature points. Thus, the use of these maps for high level purposes, such as delivery tasks or interactions with humans, requires an anchoring technique to link point-based landmarks with their associated semantics. Alternatively, in this work, objects are recognized using SIFT and then features of recognized objects are registered as landmarks. Therefore, we can create the semantic-map supporting the location of main objects in the environment without any additional anchoring process.

In most SLAM methods using Kalman filters, odometry-based robot kinematics or dynamics are used as the process model. However, those performances are very poor because of systematic factors such as the difference of wheel diameter and non-systematic factors such as slip. Moreover, in the case that a wheel encoder does not exist, such as for a humanoid robot, odometry-based methods are difficult to apply.

Therefore, to resolve such drawbacks, we propose the extended Kalman filter-based SLAM that integrates PLK optical flow and LM algorithm-based relative localization with object landmarks.

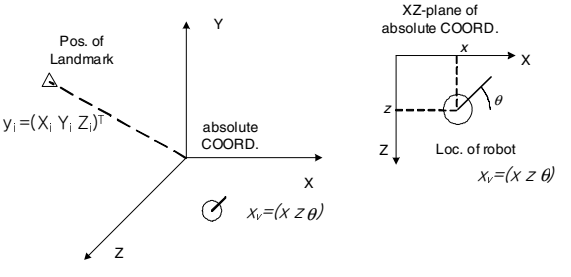

Fig. 2. System coordinate for the location of the robot x_v and the position of the landmark y_i

The State Vector and Covariance. We assume that the robot is located on a plane as in Fig. 2, so position and orientation of the robot is represented by $x_v = (\,x\ z\ \theta\,)^T$. The position of each landmark is denoted as $y_i = (\,X_i\ Y_i\ Z_i\,)^T$. Here, SIFT keypoints of objects recognized by the three dimensional OFM are represented as absolute coordinates.

To regulate uncertainty of landmarks and relationships among them we use the system state vector and covariance model proposed by [1]. Thus, we can represent the state vector p and covariance matrix Σ as

$$p = \begin{pmatrix} x_v \\ y_1 \\ y_2 \\ \vdots \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma^2_{x_v x_v} & \sigma^2_{x_v y_1} & \sigma^2_{x_v y_2} & \cdots \\ \sigma^2_{y_1 x_v} & \sigma^2_{y_1 y_1} & \sigma^2_{y_1 y_2} & \cdots \\ \sigma^2_{y_2 x_v} & \sigma^2_{y_2 y_1} & \sigma^2_{y_2 y_2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{1}$$

Three Dimensional Object Feature Model. The three dimensional object feature model(3D OFM) is defined as the model which is composed of the three dimensional SIFT keypoints extracted from object images. To make the three dimensional object feature model, we rotate and object by 20 degrees with respect to the center of gravity and then take the images using the calibrated stereo camera. This process is repeated to get 18 views images. SIFT keypoints are extracted from each image and are given three dimensional coordinates calculated with the stereo vision technique. SIFT keypoints having three dimensional coordinate are called as the three dimensional SIFT keypoints. The three dimensional SIFT keypoints in objects recognized by using 3D OFM are used as landmarks.

Process Model. In EKF, we define the process model as the estimation of current location and its covariance for the robot and landmarks referring to the state vector and its covariance matrix for the previous system. We apply the displacement $\Delta x, \Delta z, \Delta\theta$ obtained by PLK optical flow and LM algorithm-based relative localization to the process model. Let the process model of our work be given as

$$x_v(k+1|k) = f(x_v(k|k), u(k)) = \begin{bmatrix} x \\ z \\ \theta \end{bmatrix} + \begin{bmatrix} \Delta x_r \\ \Delta z_r \\ \Delta\theta \end{bmatrix}. \tag{2}$$

In (2), Δx_r and Δz_r are given as

$$\begin{bmatrix} \Delta x_r \\ \Delta z_r \end{bmatrix} = \begin{bmatrix} cos(\theta + \Delta\theta) & sin(\theta + \Delta\theta) \\ -sin(\theta + \Delta\theta) & cos(\theta + \Delta\theta) \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta z \end{bmatrix}. \tag{3}$$

The current location for the robot $x_v(k+1|k)$ is estimated by function f as shown in (2). $x_v(k|k)$ and $u(k)$ represent displacement of the previous location and motion of the robot respectively. Δx_r and Δz_r are the displacement which are transformed to the reference coordinate system of the map from the estimated displacement Δx and Δz referring to the camera coordinate system of the previous location.

$$y_i(k+1|k) = y_i(k|k), \forall i. \tag{4}$$

In (4), $y_i(k+1|k)$ represents the estimated current position of the i-th landmark, and we assume that landmarks are fixed.

The covariance for the system $\sigma^2_{x_v x_v}, \sigma^2_{x_v y_i}$ and $\sigma^2_{y_i y_j}$ are obtained as

$$\begin{gathered} \sigma^2_{x_v x_v}(k+1|k) = \nabla_{x_v} f \sigma^2_{x_v x_v}(k|k) \nabla_{x_v} f^T + \nabla_u f \sigma^2_u(k|k) \nabla_u f^T, \\ \sigma^2_{x_v y_i}(k+1|k) = \nabla_{x_v} f \sigma^2_{x_v y_i}(k|k), \\ \sigma^2_{y_i y_j}(k+1|k) = \sigma^2_{y_i y_j}(k|k), \end{gathered} \tag{5}$$

where $\nabla_{x_v} f$ and $\nabla_u f$ represent the Jacobian of the estimation function f for the state vector of the robot and the displacement respectively. These are given as

$$
\begin{aligned}
&\nabla_{x_v} f = \left[\, \tfrac{\partial f}{\partial x} \;\; \tfrac{\partial f}{\partial z} \;\; \tfrac{\partial f}{\partial \theta} \,\right], \quad \nabla_u f = \left[\, \tfrac{\partial f}{\partial \Delta x} \;\; \tfrac{\partial f}{\partial \Delta z} \;\; \tfrac{\partial f}{\partial \Delta \theta} \,\right], \\
&where \\
&\partial f/\partial x = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T, \quad \partial f/\partial z = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T, \\
&\tfrac{\partial f}{\partial \theta} = \begin{bmatrix} -sin(\theta+\Delta\theta)\Delta x + cos(\theta+\Delta\theta)\Delta z \\ -cos(\theta+\Delta\theta)\Delta x - sin(\theta+\Delta\theta)\Delta z \\ 1 \end{bmatrix}, \\
&\partial f/\partial \Delta x = \begin{bmatrix} cos(\theta+\Delta\theta) & -sin(\theta+\Delta\theta) & 0 \end{bmatrix}^T, \\
&\partial f/\partial \Delta z = \begin{bmatrix} sin(\theta+\Delta\theta) & cos(\theta+\Delta\theta) & 0 \end{bmatrix}^T, \\
&\tfrac{\partial f}{\partial \Delta \theta} = \begin{bmatrix} -sin(\theta+\Delta\theta)\Delta x + cos(\theta+\Delta\theta)\Delta z \\ -cos(\theta+\Delta\theta)\Delta x - sin(\theta+\Delta\theta)\Delta z \\ 1 \end{bmatrix}.
\end{aligned} \tag{6}
$$

Here, $(\,\Delta x \;\Delta z \;\Delta\theta\,)$ is the movement of the robot estimated by the LM algorithm.

In (5), σ_u^2 is the covariance matrix due to the noise in the process of camera motion estimation.

Measurement Model. Let the measurement model be given as $h_i(k+1) = h(y_i, x(k+1|k))$. Observation for landmark is the three dimensional coordinate of landmarks based on the calibrated stereo camera coordinate system. Measurement prediction model of h_i is given as

$$
\begin{aligned}
&h_i(k+1) = [R][y_i - x'_v], \\
&where \quad x'_v = (x \; 0 \; z)^T, R = \begin{bmatrix} cos\theta & 0 & -sin\theta \\ 0 & 1 & 0 \\ sin\theta & 0 & cos\theta \end{bmatrix}.
\end{aligned} \tag{7}
$$

Three dimensional absolute coordinate of landmark is denoted as y_i. And h_i is the function to transform absolute coordinate to camera coordinate system. x'_v represents planar coordinate of the robot based on the absolute coordinate system. Rotation between the absolute coordinate system to the camera coordinate system is denoted by R. θ in matrix R denotes the orientation of the robot.

Observation and Matching. Matching between landmarks and observed three dimensional coordinate of SIFT $z_j = [X_j \; Y_j \; Z_j\,]$ is accomplished through SIFT matching algorithm. To determine searching area for matching, we can use the innovation matrix and its covariance.

The innovation matrix $v_{ij}(k+1)$ between the predicted measurement $h_i(k+1|k)$ for landmark y_i and observation z_j is given as

$$
\begin{aligned}
&v_{ij}(k+1) = [z_j(k+1) - h(y_i, p(k+1|k))]\,, \\
&\sigma^2_{IN,ij} = \nabla_{x_v} h_i \cdot \sigma^2_{x_v x_v} \cdot \nabla_{x_v} h_i^T + \nabla_{x_v} h_i \cdot \sigma^2_{x_v y_i} \cdot \nabla_{y_i} h_i^T \\
&\qquad + \nabla_{y_i} h_i \cdot \sigma^2_{y_i x_v} \cdot \nabla_{x_v} h_i^T + \nabla_{y_i} h_i \cdot \sigma^2_{y_i y_i} \cdot \nabla_{y_i} h_i^T + \sigma^2_{R,i},
\end{aligned} \tag{8}
$$

where $\sigma^2_{R,i}$ represents the covariance of the measurement.

Searching area is determined with the Mahalanobis distance and threshold constant g^2 such as

$$v_{ij}^T(k+1) \cdot \sigma_{IN,ij}^{-2} \cdot v_{ij}(k+1) \leq g^2 \tag{9}$$

Using SIFT matching algorithm, matching is accomplished between the landmark y_i and the observation satisfying the criterion shown in (9).

Estimation of the System State Vector and Covariance. Kalman gain K can be calculated as

$$K = \sigma^2 \nabla_{x_v} h_i^T \cdot \sigma_{IN_i}^{-2} = \begin{pmatrix} \sigma_{x_v x_v}^2 \\ \sigma_{y_i x_v}^2 \\ \vdots \end{pmatrix} \frac{\partial h_i^T}{\partial x_v} \sigma_{IN_i}^{-2} + \begin{pmatrix} \sigma_{x_v y_i}^2 \\ \sigma_{y_i y_i}^2 \\ \vdots \end{pmatrix} \frac{\partial h_i^T}{\partial y_i} \sigma_{IN_i}^{-2}, \tag{10}$$

where $\sigma_{x_v x_v}^2$, $\sigma_{x_v y_i}^2$ and $\sigma_{y_i y_i}^2$ are 3×3 blocks of the current state covariance matrix Σ. $\sigma_{IN_i}^2$ is the scalar innovation variance of y_i.

The updated system state and its covariance can be computed as

$$\begin{aligned} p(k+1|k+1) &= p(k+1|k) + K \cdot v_i(k+1), \\ \sigma^2(k+1|k+1) &= \sigma^2(k+1|k) - K \cdot \sigma_{IN_i}^2(k+1) \cdot K^T. \end{aligned} \tag{11}$$

This update is carried out sequentially for each innovation of the measurement.

Registration and Deletion of Landmarks. When new three dimensional SIFT feature points are found on a recognized object using 3D OFM, absolute coordinates of those are computed using $y_n(x_v, h_n)$ of the measurement model. The computed absolute coordinates are added into the system state vector as in [1];

$$p_{new} = \begin{pmatrix} x_v & y_1 & y_2 & \cdots & y_n \end{pmatrix}^T. \tag{12}$$

Updated covariance of the system state is given as

$$\Sigma_{new} = \begin{pmatrix} \sigma_{x_v x_v}^2 & \sigma_{x_v y_1}^2 & \cdots & \sigma_{x_v x_v}^2 \frac{\partial y_n}{\partial x_v}^T \\ \sigma_{y_1 x_v}^2 & \sigma_{y_1 y_1}^2 & \cdots & \sigma_{y_1 x_v}^2 \frac{\partial y_n}{\partial x_v}^T \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_n}{\partial x_v} \sigma_{x_v x_v}^2 & \frac{\partial y_n}{\partial x_v} \sigma_{x_v y_1}^2 & \cdots & N \end{pmatrix}, \tag{13}$$

$$where \quad N = \frac{\partial y_n}{\partial x_v} \sigma_{x_v x_v}^2 \frac{\partial y_n}{\partial x_v}^T + \frac{\partial y_n}{\partial h_n} \sigma_R^2 \frac{\partial y_n}{\partial h_n}^T.$$

We can delete landmarks by deleting all rows and columns related to target landmarks from the covariance matrix.

3 Experimental Results

3.1 Performance Evaluation for Vision-Based Relative Localization

To evaluate the performance of the vision-based relative localization proposed by this work, we compare the movement of the robot with the correct answer. For the localization experiment, the robot moves 900 millimeters straight forward. Calculation interval of the robot location is 50 millimeters. Table 1 shows the error between estimation using our algorithm and the real value.

Table 1. Error of the vision-based relative localization

	PLK optical flow and LM algorithm		
	Maximum	Minimum	Average
X(millimeter)	8.785	0.122	3.310
Z(millimeter)	5.153	0.880	3.254
θ(degree)	0.969	0.013	0.366

3.2 Performance Evaluation for Object Landmark-Based SLAM Compensation Using the Extended Kalman Filter-Based Localization

It is hard to register a landmark in real-time by the object recognition. Thus, we have to do localization and semantic-map building asynchronously.

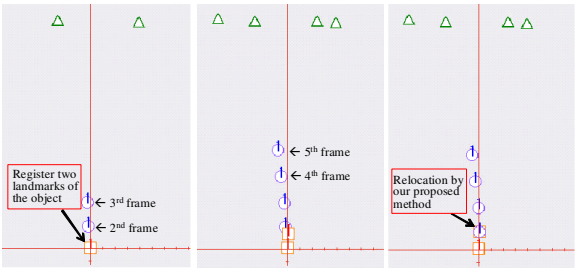

Fig. 3. The compensation process using extended Kalman filter

In Fig. 3, a circle and a rectangle indicate location of robot to be estimated by vision-based relative localization and our proposed EKF-based localization respectively. After the vision-based relative localization process, objects are recognized in a 3rd frame as shown in Fig. 3(left), and then SIFT features of the objects are registered as landmarks. On the other hand, robot localizations for 4th and 5th frames are estimated by the vision-based relative localization in real-time as shown in Fig. 3(center), while updating location of robot for 2nd frame by EKF-based localization. Fig. 3(right) represents the performances of relocation by our proposed EKF-based localization.

It is observed from Table 2 that performances of location by our proposed EKF is much better than those of vision-based relative localization.

Table 2. Error compensation result using EKF-based localization

	x (mm)	z (mm)	θ (degree)
True	0	900	0
Vision-based	-30.662	958.572	-2.383
EKF-based	3.173	911.351	-0.472

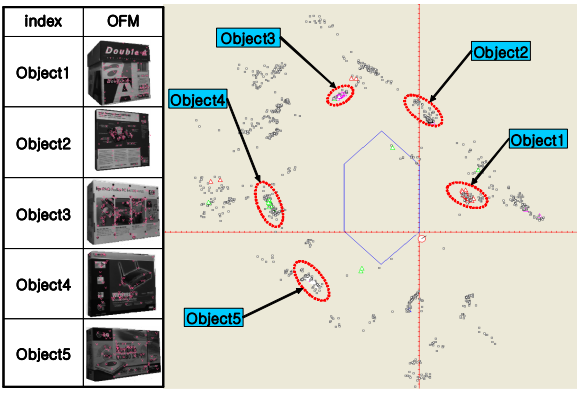

Fig. 4. Semantic-map built by robot

Semantic-Map Building. For our experiment, five objects are arranged as shown in Fig. 4. And, 3D SIFT OFMs for the five objects are also shown in Fig. 4. Keypoints of objects are recognized in the scene by using 3D SIFT OFM and are registered as landmarks with associated symbols to represent each object as shown in Fig. 4. Semantic-map indicates positions of objects as symbols associated with each object. In addition, estimated regions of each object is shown as ellipses in the semantic-map. Figure 4 shows the semantic-map formed by the robot while moving.

It is noted that object regions are informed with map-building for environment simultaneously. In traditional SLAM algorithm, however, additional process such as object recognitions and their anchorings must be required to complete this work.

Simultaneous semantic-map building with the robot localization can support new ability of the robot. First, the robot is able to simultaneously build up object-based semantic-map while carrying out delivery tasks in an unknown environment. Second, Semantic-map can be used to establish environmental ontology to be required in path planning or task planning.

4 Concluding Remarks

In this work, we have proposed autonomous semantic-map building combined with vision-based robot localization. We have used nonsymbolic SIFT-based

object features with symbolic object names as landmarks and we have used vision-based relative localization as the process model of our EKF. Thus our method is able to be applied successfully to environments in which encoders are not available, such as humanoid robots.

For real-time localization, we have used the Harris corner detector and PLK optical flow. From the relationship between three dimensional Harris corner points computed by stereo vision and corner points tracked by the PLK optical flow, we have been able to obtain the motion of the camera as well as the relative localization of the robot.

In most previous methods, landmarks have been composed of points without semantics. Thus, an additional anchoring technique was often required for interaction. However, in our work, symbolic object names with their 3D feature location have been used as landmarks. Using such symbolic object names as landmarks is very useful when humans interact with the robot.

Acknowledgement

This work is supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Korea Ministry of Commerce, Industry and Energy. And, all correspondences of this work should be addressed to I.H. Suh.

References

1. A.J. Davison and W. Murray: Simulatoneous Localization and Map-Building Using Active Vision. IEEE Transaction on Pattern Analysis and Machine Intelligence. (2002) 865–880
2. A.J. Davision: Real-Time Simultaneous Lacalisation and Mapping with a Single Camera. Proceedings of Ninth IEEE International Conference on Computer Vision. (2003) 1403–1410
3. M.W.M. G. Dissanayake, P. Newman, S. Clark, and H. F. Durrant-Whyte: A Solution to the Simultaneous Localization and Map Building (SLAM) Problem. IEEE Transaction on Robotics and Automation. (2001) 229–241
4. N. Doh, H. Choset, and W.K. Chung: Accurate Relative Localization Using Odometry. Proceedings of IEEE International Conference on Robotics and Automation. (2003) 1606–1612
5. C. Harris and M.J. Stephens: A Combined Corner and Edge Detector. Alvey Vision Conference. (1988) 147–152
6. D.G. Lowe: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. (2004) 91–110
7. B. Lucas and T. Kanade: An Interative Image Registration Technique with an Application To Stereo Vision. Proceedings of DARPA IU Workshop. (1981) 121–130
8. P.S. Maybeck: The Kalman Filter:An Introduction to Concepts. Spinger-Verlag. (1990)
9. S. Se, D.G. Lowe, and J. Little: Mobile Robot Localization And Mapping with Uncertainty using Scale-Invariant Visual Landmarks. International Journal of Robotics Research. (2002) 735–758