Design of a Simultaneous Mobile Robot Localization and Spatial Context Recognition System

Seungdo Jeong[1], Jonglyul Chung[1], Sanghoon Lee[1],
Il Hong Suh[2], and Byunguk Choi[2]

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

Abstract. In this work, we propose a simultaneous mobile robot localization and spatial context recognition system. The Harris corner detector and pyramid Lucas-Kanade optical flow are combined for robot localization. And, SIFT keypoints and its descriptors for the model-based object recognition and stereo vision technique are applied to spatial context recognition. The effectiveness of our proposed method is verified by experiments.

1 Introduction

While navigating in an environment to complete a delivery task, a mobile robot has to be able to recognize where it is, what the main objects in the scene are, and how main objects are spatially organized in the environment. Traditionally, place recognition, object recognition, and findings of spatial relations are considered separate problems. SLAM(Simultaneous Localization And Mapping) can be regarded as a popular technique to localize the mobile robot accurately and, simultaneously, to build a map of environment. To achieve SLAM, there are several different types of sensor modalities including laser range finder, sonar and monocular and/or stereo vision. A laser scanner is active, accurate, but slow and expensive, whereas sonar is fast and cheap, but usually very crude. Vision systems are passive and of high resolution.

Monocular vision-based SLAM technology as in [1] is highly desirable for a wide range of applications. However, the two dimensional vision-based SLAM technology in [1] cannot deal with object recognition and cannot find spatial relation of the objects.

Three dimensional vision approaches have been proposed in [2, 3]. Harris three-dimensional vision system DROID uses the visual motion of image corner features for 3D reconstruction [2]. Kalman filters are used for tracking features, and from the locations of the tracked image features, DROID determines both

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 945–952, 2005.
© Springer-Verlag Berlin Heidelberg 2005

the camera motion and the 3D positions of the features. Ego-motion determination by matching image features is generally very accurate in the short to medium term. However, in a long image sequence, long-term drift can occur as no map is created.

In [3], Se describes a vision-based mobile robot localization and mapping algorithm, which uses Scale Invariant Feature Transformation(SIFT) keypoints as scale-invariant image features and natural landmarks in unmodified environments. The invariance of these features to image translation, scaling and rotation makes them suitable landmarks for mobile robot localization and map building. Feature viewpoint variation and occlusion are taken into account by maintaining a view direction for each landmark.

However, the SIFT stereo approach suffers from computational complexity, which can require 2-3 seconds of processing time per SIFT stereo image on contemporary PC hardware. With a robot velocity of 300mm/sec, this could result in a separation of up to 1m distance between consecutive SIFT stereo image locations. In this case, SIFT keypoints of the previous and current locations may be different, and exact point-to-point correspondence matching may not be possible. This would incur a 50% or greater error in relative position accuracy. In addition to the above difficulty, all these vision approaches are not concerned with recognition of objects and their spatial relations.

It is remarked that spatial relation is a key contextual information. To understand a local context, it is essential to firstly recognize objects within the local environment, and then to find out their spatial relations.

In this work, SIFT keypoints are used for model-based object recognition. And, object model includes three dimensional shape information. Once objects are detected, then Harris corner stereos are used to estimate spatial relations. In contrast to SIFT stereos, Harris corners can be calculated in a few hundred milliseconds on contemporary PCs, which can support 3D object motion tracking in a real-time. To ensure correct corner-to-corner correspondence between stereo images taken at consecutive locations, Lucas-Kanade optical flows are applied. This technique makes Harris corner detection less sensitive to changes in scale. Ego-motion of the mobile robot is then estimated by least square minimization of Harris corners of the matched landmark objects. After all, 3D relation between the mobile robot and main objects can be found, and thus 3D spatial relations between objects with respect to the robot can be also obtained. As a result, a mobile robot is able to recognize where it is, what the main objects in the scene are, how main objects are spatially organized in the indoor environment of the real world. Therefore, a mobile robot is able to complete the delivery task.

2 Overview of Simultaneous-Localization-and-Spatial-Context-Recognition System

Consider block diagram of our proposed simultaneous localization and spatial context understanding system as shown in Fig. 1. In Fig. 1, proposed system is composed of two functional parts. The robot localization module is designed to work with a sampling period of hundreds milliseconds. First of all, feature points are extracted by corner detection module within the images taken from the stereo camera. Next, the feature tracking module tracks the extracted feature points. These tracked feature points are used as point-based landmarks for the robot localization. The landmarks which is used for the robot localization are represented as 3D coordinates by using stereo matching. The robot localization module estimates the robot location by using the relationships between 3D landmarks and corresponding points projected to the current image.

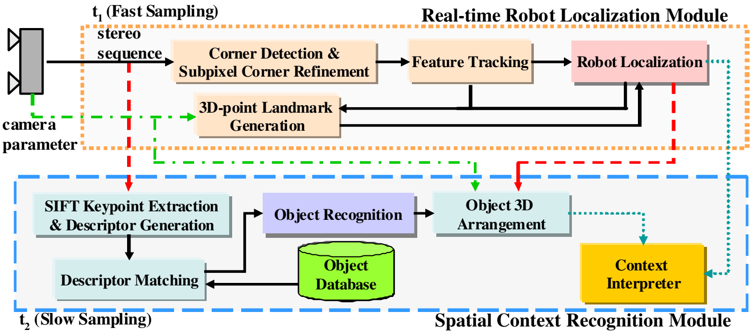


Fig. 1. Block Diagram of Simultaneous Localization and Spatial Context Understanding System

To understand the spatial context, the object recognition is essential. Our spatial context understanding module as in Fig. 1 is composed of the feature extraction module, the object recognition module, and the interpreter for spatial arrangement of the recognized objects. The context understanding module does not require real-time processing because there is no difficulty to understand the arrangement of the main objects and the spatial structure of local environment, even though this module does not operate at each frame. Therefore, we organize to work with slow sampling period contrary to the robot localization module, and to recognize the main objects exactly.

3 Design of a Real-Time Robot Localization Module

3.1 Detection of Harris Corners

The SIFT has been highlighted in the robot vision community recently, which is a method to extract and describe the feature point. SIFT uses Difference-of-Gaussian(DoG). Maxima and minima of the DoG images are detected by comparing a pixel to its 26 neighbors in 3×3 regions at the current and adjacent scales. These maxima and minima become the keypoints. In addition, the Gaussian image is down-sampled by a factor of 2, and the process is repeated. Therefore, the SIFT features are invariant to scale. A keypoint descriptor is created by first computing the gradient magnitude and orientation at each image sample point in the region around the keypoint location. To achieve orientation invariance, the coordinates of the descriptor and the gradient orientations are rotated relative to the keypoint orientation. Thus, SIFT is invariant to image rotation and scale, and robust across a substantial range of affine distortion, addition of noise, and change in illumination [4].

However, the algorithm to estimate the exact position of SIFT keypoint requires heavy computational cost. And, the position of a keypoint is not consistent. Thus, SIFT may not be a good feature candidate for a real time localization.

Alternatively, in this work, we will use Harris corner detector to extract the feature points in a real-time. The Harris corner detector [5, 6] is based on the autocorrelation function. To be specific, let $A(x, y)$ be defined by

$$A(x,y) = \begin{bmatrix} \sum(I_x(x_k,y_k))^2 & \sum I_x(x_k,y_k)I_y(x_k,y_k) \\ \sum I_x(x_k,y_k)I_y(x_k,y_k) & \sum(I_x(x_k,y_k))^2 \end{bmatrix}, \tag{1}$$

where (x_k, y_k) are the points in an window centered on (x, y). Then, corner points are detected if the autocorrelation matrix A has two significant eigenvalues.

In order to verify the Harris corner is more suitable for real time localization than SIFT, Table 1 shows the performance comparison of the Harris corner detector and the SIFT algorithm. The Harris corner detector includes the corner extraction and the sub-pixel refinement, whereas, the SIFT algorithm includes the extraction of keypoints and the generation of descriptors. Although, the number of feature points of the Harris corner detector and the SIFT algorithm looks similar, the Harris corner detector has relatively low computational cost when compared to the SIFT algorithm. Thus, Harris corner detector is more suitable for the real-time processing than SIFT. In Table 1, the unit of computation time is millisecond.

Table 1. Feature extraction time per image

Image	Harris corner detector				SIFT algorithm			
size	Feature Points	Max	Min	Average	Feature Points	Max	Min	Average
176 × 144	252.8	45	14	27.65	270.4	860	406	651.24
320 × 240	474.1	47	14	28.86	492.1	3157	1875	2292.57
640 × 480	696.4	125	31	62.35	715.7	8891	6469	7660.8

3.2 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. In other words, the motion of neighborhood points about a corner is almost same. Supposing the consecutive frames called H and I, the brightness of the corresponding point is the same. That is, the motion of a pixel is represented as

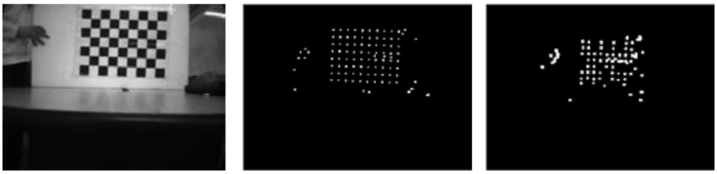
$$H(x, y) = I(x + u, y + v). \tag{2}$$

With the first-order Taylor expansion of $I(x + u, y + v)$, we can derive the final optical flow equation as

$$I_t + \nabla I \cdot [u\ v] = 0. \tag{3}$$

If we estimate the motion of one pixel only, there will be the aperture problem. To resolve this aperture problem, we can apply the local smoothness constraint; the motions of neighborhood pixels are almost same. Therefore, the matrix form of the optical flow equation should be considered to include $n \times n$ window around the feature point.

The Lucas-Kanade optical flow is suitable to the small motion of one pixel difference [7]. However, we can hardly expect such a small motion as in the practice. Thus, we apply the pyramid Lucas-Kanade optical flow. The pyramid Lucas-Kanade optical flow first requires to generate the Gaussian image pyramid. Then optical flow is estimated from the lowest level of sub-sampled image to original image iteratively by using image warping and up-sampling process.



(a) Input Image (b) Harris Corner (c) SIFT

Fig. 2. Comparison of the Harris corner and the SIFT stereo

The experimental result for stereo matching using the Harris corner detector following the pyramid Lucas-Kanade optical flow and using the SIFT algorithm is shown in Fig. 2. Our algorithm not only extracts the feature points with coherency, but also tracks the feature points exactly as shown in Fig. 2. However, the SIFT algorithm has little coherency though the extracted feature points are invariant to the variation of image. This shows that there exists a difference in the repeatability in the sense that the same feature point is extracted consecutively.

3.3 Robot Localization

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

$$\begin{bmatrix} x \\ y \\ w \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \left[R\,|t \right] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{4}$$

where R is 3×3 rotation matrix and t is 3×1 translation matrix.

Projection matrix is composed of two parts. One is the intrinsic parameter matrix which includes internal information of camera. This matrix represents the relationship between camera coordinate system and image coordinate system. In intrinsic parameter matrix, f_x, f_y represent the focal length for each direction respectively, (c_x, c_y) represents the principal point. Another 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 [8]. With that the intrinsic parameter matrix is already obtained by in processing of the stereo camera calibration. Therefore, if the projection matrix is estimated with coordinates of corresponding points in 3D space and in 2D image, we can obtain the motion of camera.

It is the non-linear estimation problem to estimate the projection matrix with numerous corresponding points. In this work, we estimate rotation and translation components by using Levenberg Marquardt least square minimization method [9]. The robot localization is accomplished by converting camera motion to the 3D space coordinate system.

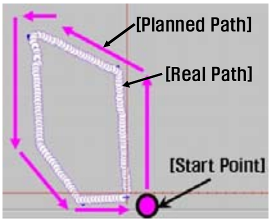


Fig. 3. Robot localization result

In order to verify the localization performance for the successful delivery task, we observe how precise the robot recognizes location and it returns to the start position exactly with the planed path. Fig. 3 shows the localization result by using our localization module for the planned path. The robot moves from start position and return to start position navigating about 6.5 meter including with moving in front and rotating. Last position error is about 10 centimeter by using the proposed localization method. Each error of moving robot position is shown in the Table 2. Each numerical value is surveyed per 50 millimeters.

4 Design of Spatial Context Recognition Module

4.1 Object Recognition

It is required to extract the precise feature points more than the Harris corner for object recognition. In this work, the SIFT keypoints and their descriptors are used for model-based object recognition [10]. The object database is previously builded in such a way that this database includes the information of the

Table 2. Error of the Robot Localization

	Harris Corner + Pyramid L-K		
	Maximum	Minimum	Average
X(millimeter)	17.72	0.52	4.896
Y(millimeter)	5.42	0.14	3.12
Z(millimeter)	18.22	1.62	6.146
θ(degree)	0.418	0.021	0.071

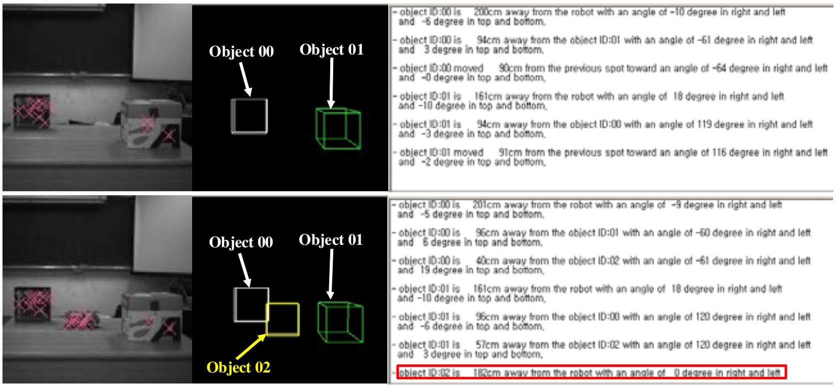
SIFT descriptor for each image patch. The object recognition is accomplished by matching between the SIFT keypoints extracted from current scene with the SIFT keypoints of each object in the object database.

4.2 Extraction of Spatial Relations of Objects

The recognized object has only 2D coordinate in current scene yet. So, to represent relative arrangement of each object in the 3D space we need to know the 3D coordinate of each object. The disparities of the SIFT keypoints within each object are calculated individually and then, the 3D coordinate of each keypoint is obtained by using these disparities. Because the 3D coordinate of each keypoint is already classified into each object, the region including all 3D keypoints of object is the 3D location of the object. However, this location refers to the current camera. So, we need to know the location of the current camera in order to obtain the position of the object in the absolute space. The current camera position can be acquired from the robot localization module in the real-time processing sub system. Therefore, we can finally represent the 3D coordinate of respective object with respect to the absolute coordinate system using the robot localization information and 3D location of each object with respect to the current camera. Locations of objects interpreted by interpreter module are informed to robot localization module to correctly compensate the cumulated error caused by long-term operations of the Harris corner detector and pyramid Lucas-Kanade optical flow. Fig. 4 shows an example of recognition of spatial relations of objects. This example shows that the interpreter module classifies new object and exactly interprets its spatial relation with respect to previous objects, whenever it appears newly.

5 Conclusion

In this work, we have proposed a method to not only localize a mobile robot but also recognize the spatial context. This was implemented by effective hybridization of several paradigms; for localization, Harris corner detector and pyramid Lucas-Kanade optical flow. And for recognition of spatial relations, 2D and 3D SIFT keypoints were employed. Our proposed method was successfully applied to a mobile robot navigation.



```
- object ID:00 is    200cm away from the robot with an angle of -10 degree in right and left
  and  -6 degree in top and bottom.

- object ID:00 is     94cm away from the object ID:01 with an angle of -61 degree in right and left
  and   3 degree in top and bottom.

- object ID:00 moved     90cm from the previous spot toward an angle of -64 degree in right and left
  and  -0 degree in top and bottom.

- object ID:01 is    161cm away from the robot with an angle of  18 degree in right and left
  and -10 degree in top and bottom.

- object ID:01 is     94cm away from the object ID:00 with an angle of 119 degree in right and left
  and  -3 degree in top and bottom.

- object ID:01 moved     91cm from the previous spot toward an angle of 116 degree in right and left
  and  -2 degree in top and bottom.
```

```
- object ID:00 is    201cm away from the robot with an angle of  -9 degree in right and left
  and  -5 degree in top and bottom.

- object ID:00 is     96cm away from the object ID:01 with an angle of -60 degree in right and left
  and   6 degree in top and bottom.

- object ID:00 is     40cm away from the object ID:02 with an angle of -61 degree in right and left
  and  19 degree in top and bottom.

- object ID:01 is    161cm away from the robot with an angle of  18 degree in right and left
  and -10 degree in top and bottom.

- object ID:01 is     96cm away from the object ID:00 with an angle of 120 degree in right and left
  and  -6 degree in top and bottom.

- object ID:01 is     57cm away from the object ID:02 with an angle of 120 degree in right and left
  and   3 degree in top and bottom.

- object ID:02 is    182cm away from the robot with an angle of   0 degree in right and left
```

Fig. 4. Spatial relations of objects

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.

References

1. A. J. Davison: Real-Time Simultaneous Localization and Mapping with a Single Camera. Proceedings of International Conference on Computer Vision. (2003) 1403–1411
2. C. Harris: Geometry from visual motion. Active Vision. MIT Press. (1992) 264–284
3. S. Se, D. Lowe, and J. Little: Mobile Robot Localization and Mapping with Uncertainty using Scale-Invariant Visual Landmarks. International Journal of Robotics Research. (2002) 735–758
4. D. Lowe: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. Vol. 60 (2004) 91–110
5. C. Harris and M. J. Stephens: A combined corner and edge detector. Alvey Vision Conference. (1988) 147–152
6. C. Schmid, R. Mohr, and C. Bauckhage: Evaluation of Interest Point Detectors. International Journal of Computer Vision. Vol. 37, No. 2 (2000) 151–172
7. B. Lucas and T. Kanade: An interative image registration technique with an application to stereo vision. Proc. DARPA IU Workshop. (1981) 121–130
8. R. Hartley and A. Zisserman: Multiple View Geometry in Computer Vision. Cambridge University Press. (2000)
9. W. Press, S. Teukolsky, W. Vetterling, and B. Flannery: Numerical Recipes in C++. Cambridge University Press. (2002)
10. S. Helmer and D. Lowe: Object Class Recognition with Many Local Features. Workshop on Generative Model Based Vision. (2004)