2005 IEEE/RSJ International Conference on Intelligent Robots and Systems

Recognition-based Indoor Topological Navigation Using Robust Invariant Features

Zhe Lin, Sungho Kim and In So Kweon

Department of Electrical Engineering and Computer Science
Korea Advanced Institute of Science and Technology
373-1, Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea
{limcher, shkim}@rcv.kaist.ac.kr, iskweon@kaist.ac.kr

Abstract **– In this paper, we present a recognition-based autonomous navigation system for mobile robots. The system is based on our previously proposed Robust Invariant Feature (RIF) detector. This detector extracts highly robust and repeatable features based on the key idea of tracking multi-scale interest points and selecting unique representative local structures with the strongest response in both spatial and scale domains. Weighted Zernike moments are used as the feature descriptor and applied to the place recognition. The navigation system is composed of on-line and off-line two stages. In the off-line learning stage, we train the robot in its workspace by just taking several images of representative places as landmarks. Then, in the on-line navigation stage, the robot recognizes scenes, obtains robust feature correspondences, and navigates the environment autonomously using the Iterative Pose Converging (IPC) algorithm which is based on the idea of the visual servoing technique. The experimental results and the performance evaluation show that the proposed navigation system can achieve excellent performance in complex indoor environments.**

Index Terms – Object recognition, autonomous navigation, path planning, iterative pose converging, visual servoing.

I. INTRODUCTION

The use of local features in the context of object/place recognition and vision-based robot localization and mapping has been successful due to their invariance and power to handle occlusions and background clutters. Many local feature detectors have been proposed in the context of object recognition [1], [2], [3], [4], place recognition [9], [10], [11], [12], vision-based localization and navigation [15], [16], [17] and SLAM [13], [14]. Although many of these detectors have shown very good results and wide applicability, more reliable local feature detectors are still needed to improve the performance of visual recognition, localization and autonomous navigation. RIF detector [5] has been aimed at detecting more robust and repeatable features. It first detects multi-scale interest points and tracks them over scales to obtain the structure-wise feature representation. Next, according to the information from each group of points, shape adapted local invariant regions are extracted. Finally, rotation invariant weighted Zernike moments are calculated on the normalized image patches for the local description.

The place recognition proceeds in the following two stages. In the off-line learning stage, we detect RIFs from each model image and store them to the database. And in the on-line processing stage, we similarly detect features from the input image and compare the descriptor vectors to the stored database ones based on the nearest neighbor principle using Euclidean distance measure. Based on the initial feature correspondences, the optimal homography and the relative pose between the input and model image can be optimally estimated.

We applied the recognition and pose estimation algorithm to the single camera-based mobile robot indoor topological navigation. First, we learn the robot navigation environment or paths by taking images from the representative nodes in the map and storing corresponding RIFs to the DB. Next, through the place recognition system, the feature correspondences and relative pose can be robustly estimated. Finally, the IPC algorithm is used to iteratively correcting the robot pose by converging the current image pose to the corresponding DB image pose.

In the next section, we briefly introduce the RIF detector. Section III introduces the place recognition system and Section IV describes the proposed recognition-based path planning and autonomous navigation system. Section V presents the experimental results for the proposed single-camera-based autonomous navigation system. Finally, in Section VI, we summarize this paper.

II. FEATURE EXTRACTION

We use the RIF detector [5] for low-level image feature extraction. RIF detection process is composed of two stages of local region detection and description.

A. Local Region Detection

Given an input image, first, we incrementally smooth it with Gaussian kernel to construct the multi-scale image representation. Next, from this multi-scale representation, the second moment matrix is calculated at every pixel location in each scale level image. Then, the multi-scale interest points are localized at local peaks of the normalized Harris measure in the image domains [3], [4].

As shown in Fig. 1, after detecting multi-scale interest points, we track to group these multi-scale interest points for the local structure-wise feature representation. The idea is to cluster those multi-scale interest points corresponding to the same local structure. The linking process is gradually propagating from the highest scale level to lower scale levels based on the principle of nearest neighbor search strategy within the corresponding uncertainty ranges. The tracking and grouping process continues until no corresponding link points existing in the lower scale level within the allowed uncertainty ranges.

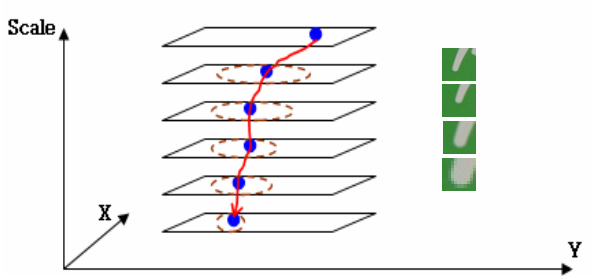

Fig. 1 Scale space interest point tracking. Planes represent scale space images, blue dots are examples of interest points belong to the same local structure, circles are uncertainty regions of the interest points.

The uncertainty range propagation is very important here, as it largely affects the grouping results. For example, when many texture components existing in the image, the relative distances between local structures can be very small, hence many false groupings can be generated, which consequently decreases the number of correct estimations. Since the interest point propagation in the scale space is directly related to the scale change rule, we approximate this interest point propagation rule as the following exponential model:

$$R(s,l) = k * s^{l} \tag{1}$$

where $R(s,l)$ represents the radii of the uncertainty regions, k is a constant factor, s is the parameter of the function and l is the scale level index. The parameter s is chosen as the same as the scale factor σ between consecutive scale levels. We have tested various values of k with respect to the resulting grouping error. Experimentally, the minimum error is obtained at $k = 0.5$ and the parameter k in (1) is fixed at this value.

Then, we use the normalized Harris measure for the scale selection. This measure naturally fits the unified framework of searching for the strongest response in both spatial and scale domains. In addition, by the use of tracking and grouping algorithm, the ambiguity and inaccuracy in scale selection can be reduced maximally. We observe the trace of normalized Harris measure responses along the LCSs and search for the local peaks in the trace. Then, we select a unique representative scale at the strongest peak point so that the corresponding region appears the *most corner-like structure*. This is similar to the visual perception on corner structures since many evidences from neuropsychology and cognitive science have shown that Human Visual System has stronger attention or response in high curvature points or corners than edges and other features. We can further estimate the exact scale by fitting parabola to each selected peak.

We have tested the performance of the RIF detector under various image variations. The Hangul image set is selected as the experimental samples. Fig. 2 shows some examples of detection results in various geometric and photometric changes. This results show that the RIFs are very consistent to large scale, viewpoint and illumination changes and robust in a scale range of 1/4 to 4 and viewing angle range of $-40^{\circ} < \theta < 40^{\circ}$.

Fig. 2 Examples of RIF detection under large scale, viewpoint and illumination changes.

B. *Local Region Description*

For feature description, we first normalize the extracted local patches to the canonical 10×10 circles. Then, dominant orientation is estimated using gradient distribution on the local image patch [4]. We assume the conventional linear illumination change model:

$$I' = sI + o \tag{2}$$

Based on this simple model, we normalize the image patches by linearly shifting its mean and variance to fixed values. In this way, the scale s and the offset o can be eliminated to get the illumination normalized patch. For the description, weighted Zernike moments are used [6]. Zernike moments have superior properties in terms of image content representation, information redundancy and noise characteristics [7] so they can be reliably used in the recognition problem. It is defined over a set of complex polynomials which form a complete orthogonal set over the unit disk. Zernike moments are calculated by projecting the image intensity onto these orthogonal basis functions. Gaussian window is used to weight the image patch before calculating the descriptors. Moreover, since the individual components are uncorrelated, Euclidean distance can be used as the similarity measure for matching. The weighted Zernike moments are calculated as follows:

$$A_{nm} = \frac{n+1}{\pi} \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} W(x,y) f(x,y) [V_{nm}(x,y)]^{*} \tag{3}$$

$$V_{nm}(x,y) = R_{nm}(x,y) \exp(jm \tan^{-1}(y/x)) \tag{4}$$

$$R_{nm}(x,y) = \sum_{s=0}^{(n-|m|)/2} (-1)^{s} \frac{(n-s)!}{s! \left(\frac{n+|m|}{2} - s\right)! \left(\frac{n-|m|}{2} - s\right)!} (x^2+y^2)^{(n-2s)/2} \tag{5}$$

where $V_{nm}(x,y)$ is an orthogonal basis function, $R_{nm}(x,y)$ is radial polynomial, $f(x,y)$ is the image function and $W(x,y)$ is the Gaussian weighting function. We calculate the 24-dim Zernike moment descriptor based on (3), (4), (5) using a fast implementation [8].

III. Place Recognition System

The framework of the place recognition system [5, 6] is shown in Fig. 3. It consists of on-line and off-line parts. In the off-line learning stage, RIFs are detected from the model images and the resulting local descriptors are stored in the database. In the on-line recognition stage, features detected from the input and model images are pushed into the searching engine to find the most similar match. The Euclidean distance is used to evaluate the similarity between descriptors. We use the Approximate Nearest Neighbors (ANN) search algorithm and a probabilistic voting technique for efficient DB indexing. The recognition verification is done by performing a 4 point RANSAC algorithm on initial feature correspondences. This stage ensures the recognition result to be correct by estimating the optimal homography and counting inliers and outliers. Consequently, we can obtain robust feature correspondences with very few false matching pairs and the relative image pose can be optimally estimated from the correspondences. Fig. 4 shows a stretch of the probabilistic voting and DB indexing process. It is composed of three parts: test image, DB images and feature descriptor space. The feature space is constructed from DB images and features from input image are voted to the DB feature space to find their nearest neighbors. Finally, based on the voting result, the maximum likelihood detection and image relative pose estimation is achieved.

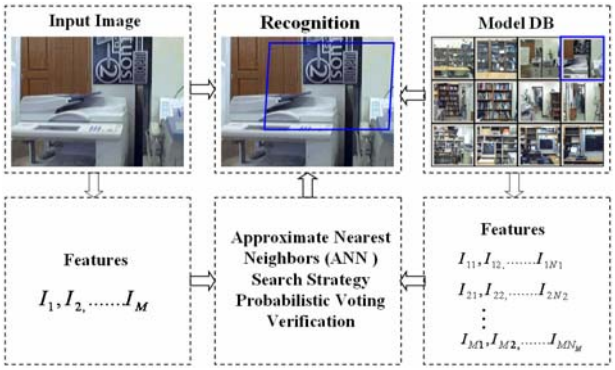

Fig. 3 Place recognition system framework. It consists of off-line learning stage (last column) and on-line recognition stage (first two columns).

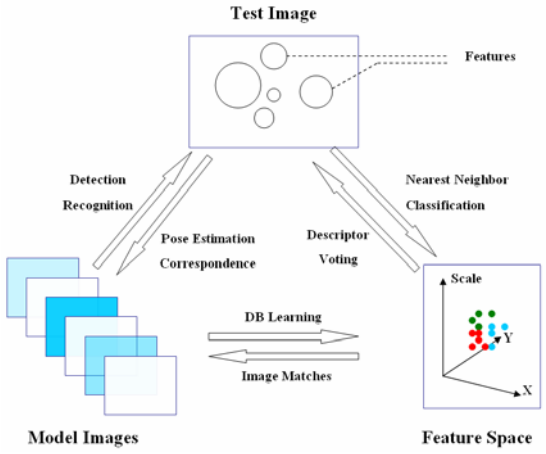

Fig. 4 Probabilistic voting and DB indexing process.

IV. Single Camera-based Autonomous Navigation

The key to the proposed autonomous navigation system is the estimated image relative pose information. Here, we use affine transformation to represent the relative image pose information. For estimating the affine transformation, we first manually drive the robot in its workspace and capture images at representative locations that the robot is expected to do critical motion. Next, through the place recognition system, robust feature correspondences can be found from the recognition process and consequently the global homography matrix can be obtained. Then, the optimal affine transformation matrix and the translation vector can be estimated by approximating the homography matrix to the simpler form of affine transformation as shown in (6).

$$H = \begin{pmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{pmatrix} -> \begin{pmatrix} A^* & t \\ 0 & 1 \end{pmatrix} \quad (6)$$

The robot motion is planned to iteratively correct its motion and converges to the optimal pose matching with database pose. The iterative pose converging is based on the estimated landmark ID and the estimated affine transformation.

We use the basic visual servoing technique to control the robot movements when performing certain tasks. It mimics the active sensing and measuring mechanism from the Human Visual System. Given the relative pose relation between the current and the recognized DB image, we can design a control rule so that the current image pose gradually converges to the matched DB image pose. For example, in Fig. 5, we can control the robot motion so that the corresponding vertex pairs in the current and the DB image poses meet each other (A'-A, B'-B, C'-C, D'-D). In our implementation, the robot motion is specifically controlled by the translation t and the optimal affine transformation A^* as these terms indicate directly the current camera pose shift.

Motion control rule:

$$t = \begin{pmatrix} t_x \\ t_y \end{pmatrix} \rightarrow \begin{cases} \text{Rotation correction} & \|t\| > \tau \\ \text{Forward} & \|t\| < \tau \ \ and \ \ \det(A^* - I) > \varepsilon \\ \text{Stop} & \|t\| < \tau \ \ and \ \ \det(A^* - I) < \varepsilon \end{cases}$$

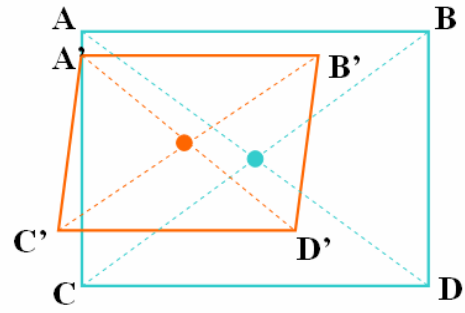

Fig. 5 An example of the pose relation between current image and matched DB image.

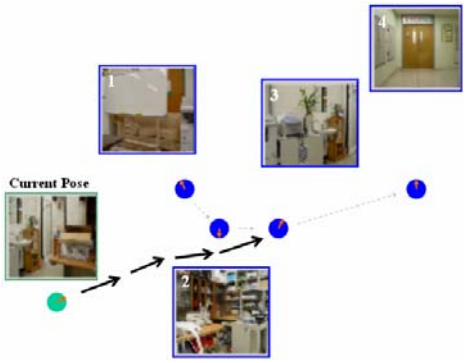

Fig. 6 Navigation scheme. Five images are one current camera input image and four place DB images, circular dots with arrows are current and DB robot poses (or landmarks), and black arrows represent the desired converging paths of the robot.

Based on the above iterative pose converging technique, the autonomous navigation is performed as follows:

1. Topologically localizing the robot using the place recognition system.
2. Obtaining robust feature correspondences.
3. Calculating relative pose between current image and the recognized DB landmark.
4. According to the relative pose information, gradually controlling the robot to iteratively converge the current pose to the DB landmark pose.
5. Adjusting the robot pose using odometry for moving to the next landmark and so forth.

Fig. 6 shows an example of the autonomous navigation scheme when the robot performs moves from a location in the lab to another lab through corridor. This example scheme shows that the robot initially arrives to the third DB landmark in the map and then moves to the next landmark.

V. Experimental Results

This section presents experimental results for the proposed place recognition and its application to single camera-based autonomous navigation.

A. Experimental Setting

Fig. 7 shows our experimental platform, consisting of a Pioneer II mobile robot and a lab-top computer with a USB camera connected to it. The camera captures 320*240 image sequences as the input to the navigation system.

Fig. 7 Experimental platform – Pioneer II robot with a single camera.

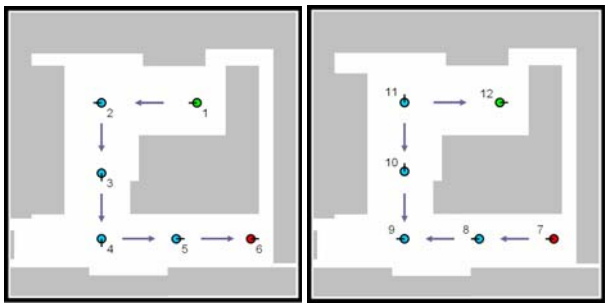

Fig. 8 Experimental map with robot poses for capturing DB place images.

Fig. 9 Place DB images taken at 12 landmark poses.

Fig. 8 shows the experimental map of a $7m \times 7m$ indoor lab environment. We use this map for the place recognition and navigation experiments. In this map, gray and white parts represent occupied regions and open spaces, respectively. We choose 12 landmark robot poses as shown in Fig. 8 and capture images in each location for storing to the DB. These poses are mostly selected at the critical locations in the environment such as corners or ends of open spaces. The landmark DB images are shown in Fig. 9.

B. Place Recognition Results

We use the images in Fig. 9 as the database for testing the place recognition performance. The testing image sequences are collected by manually drives the robot through the desired path from landmark 1 to landmark 6 and then from landmark 7 to landmark 12 indicated in Fig. 8. In the test, we totally collected 86 images continuously near the path. And in order to test the performance under illumination changes, these images are taken at the daytime and the landmark DB images are taken at night.

Fig. 10 shows some examples of the test images and their recognition results. The first two rows show recognition results under significant scale and viewpoint changes. And, the third row is a recognition result under large illumination change. The last row of the example is a recognition result under occlusion.

All these correct recognition examples show that our place recognition system is very robust under large geometric, photometric or dynamic changes and can be efficiently used in the topological localization and autonomous navigation.

Fig. 10 Experimental results of the place recognition. Examples of place recognition under various changes (scale, viewpoint, illumination and/or dynamic changes). Left: Input images, Right: Recognized landmark images.

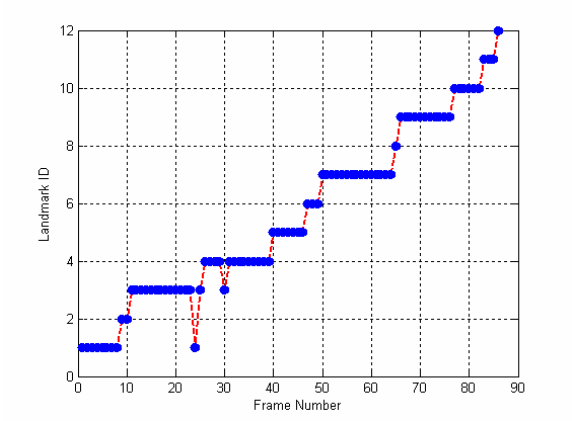

Fig. 11 Place recognition result without the previously accumulated recognition information.

Fig. 11 shows the quantitative recognition performance in the above example. The total number of correct recognition case is 86 out of 88 frames (Recognition rate: 98%) when the robot moves from the starting location (landmark 1) to the destination (landmark 6, 7) and then come back to the starting point (landmark 12) through the desired paths. False recognition results are mostly obtained in a place with very few features or in very ambiguous places. Most of the false recognition results can be reduced by incorporating the previous recognition records to the current recognition using the probabilistic Markov Chain Model in the continuous input video recognition. For example, the method in [10] shows the application of the probabilistic Markov model for improving the place recognition results in visual feature-based mobile robot topological localization.

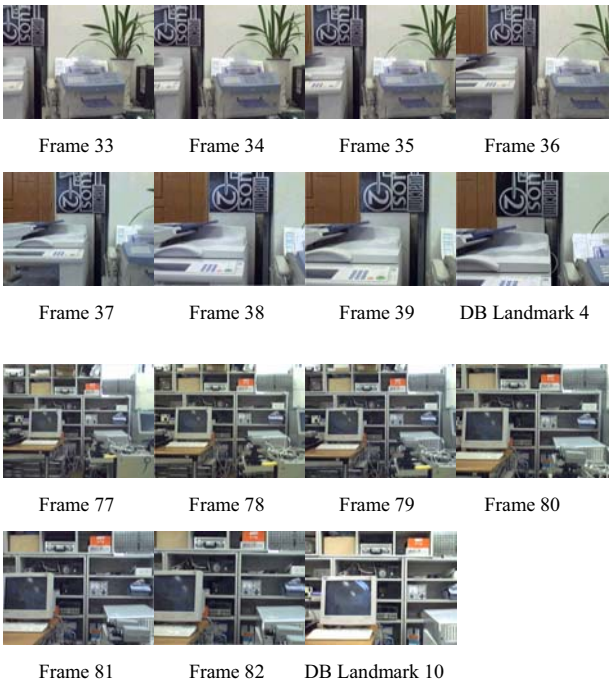

Fig. 12 Examples of iterative pose converging process.

C. *Pose Converging Results*

We have tested the pose converging performance from arbitrary poses to the DB landmark poses. The initial poses are collected near the navigation path with the orientation similar to the desired paths. As a result, in most cases, the robot pose successfully converges to the recognized DB landmark pose. The failed cases are caused by false recognition results. Because the false recognition results produce false relative image pose and control law which will diverge the robot pose eventually.

Fig. 12 shows examples of the iterative pose converging process. In the first example, the camera pose converges in 7 iterations after initially recognizes the corresponding DB landmark 4 and in the second example, the pose converges in 6 iterations after initially recognizes the DB landmark 10. The converging speed depends largely on the motion control parameters and the initial robot poses. In the chosen parameter setting, the current pose converges in average within 10 iterations.

D. *Topological Navigation Results*

The topological navigation is tested in a U-shaped indoor lab environment as shown in Fig. 8. The robot moves from the starting location to the destination and returns to the starting location. We put the robot in the starting location and test if it will correctly follow the route, arrives at the destination and returns to the starting location. The navigation paths are shown as oriented line segments between neighboring landmark locations in Fig. 9. The results show that the robot correctly achieved the goal (from landmark 1 to landmark 12 in Fig. 8) with 90% success rate among over 20 navigation trials. This high success rate is due to the robustness of the place recognition and the pose estimation. Success rate also depends largely on the number of landmarks over the paths and the failed cases mostly

occur when the robot recognizes ambiguous locations or it collides with objects in occupied regions.

For testing the kidnapping recovery performance of the system, we have put the robot in arbitrary poses near the navigation route and performed the same experiment. As a result, the robot successfully finds the nearest landmark in most cases and converges to it and continues its navigation to the next landmark.

As the further work, for estimating the exact robot pose (x, y, θ) after topological localization, we can use the pose estimation scheme in [16], [17] or [18]. Based on the robust feature correspondences and camera intrinsic parameters, we can obtain the relative robot pose to the corresponding landmark and globally estimate the robot pose up to map scale. For metric localization, the remaining 1-DOF scale factor can be obtained by using odometry information.

VI. Conclusions and Future Works

We have developed a place recognition-based autonomous navigation system for indoor mobile robots. Robust invariant features are used as the primitives of the recognition module and they are further applied to the topological localization and navigation. Based on the relative image poses between current and landmark images, the path planning is performed using a control law similar to visual servoing technique. Using IPC algorithm, the robot can gradually rectify its poses to the landmark poses and finally converges to them. This motion planning and navigation idea is fundamentally based on the model of HVS-based navigation principle. The processing time for the place recognition and path planning for one frame is approximately 0.3 seconds in a Pentium-IV 1700 processor.

For improving the place recognition performance, we are still working on the generalization of RIF detector to the affine invariance case and the development of more robust and computationally efficient local region descriptor with high distinctiveness.

For navigation, currently we assume 2D environment and the system estimate a 3-DOF robot poses. In our future work, we need to generalize it to 3D environment with full 6-DOF poses so that our system can also be applied to more general non-planar applications. Another future work is to develop an automatic map learning algorithm, odometry fusion-based accurate pose estimation algorithm between two consecutive nodes and eventually to build a fully autonomous single camera-based indoor metric-localization and navigation system.

As an application, the proposed system can be efficiently used for mobile robots providing services in a small-scale indoor environment such as home or office. Given the pre-specified map landmarks, the robot can plan its paths and navigate to the arbitrary destination. Another application is that the system can be used as the navigation guide for visually impaired people.

Acknowledgment

This work is partially supported by Ministry of Information and Communications (MIC) and NRL (code# M1-0302-00-0064) of MOST, Korea.

References

[1] D. G. Lowe. “Object recognition from local scale-invariant features,” *In Proceedings of the 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 1150–1157, 1999.

[2] D. G. Lowe. “Distinctive image features from scale invariant keypoints,” *International Journal on Computer Vision*, 60(2):91-110, 2004.

[3] K. Mikolajczyk and C. Schmid. “Indexing based on scale invariant interest points,” *In Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, pages 525–531, 2001.

[4] K. Mikolajczyk and C. Schmid. “An affine invariant interest point detector,” *In Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, volume I, pages 128–142, May 2002.

[5] Zhe Lin, Sungho Kim and In So Kweon, “Robust invariant features for object recognition and mobile robot navigation,” *In Proceedings of IAPR Conference on Machine Vision Applications, Tsukuba Science City, Japan*, 2005.

[6] S.H. Kim, I.C. Kim and I.S. Kweon. “Probabilistic model-based object recognition using local zernike moments,” *In The IAPR Workshop on Machine Vision Applications, Nara-ken New Public Hall, Nara, Japan*, Dec. 11-14, 2002.

[7] T. Drummond and R. Cipolla. “Real-time visual tracking of complex structures,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):932 –946, July 2002.

[8] Chee-Way Chong, P. Raveendran, R. Mukundan. “A comparative analysis of algorithms for fast computation of Zernike moments,” *Pattern Recognition* 36(3): 731-742, 2003.

[9] A. Torralba, K. Murphy, W. Freeman and M. Rubin, “Context-based vision system for place and object recognition,” *In Proceedings of the 9th International Conference on Computer Vision, Nice, France*, 2003.

[10] F. Li, J. Kosecka, “Vision based topological markov localization,” *In Proceedings of International Conference on Robotics and Automation, Barcelona, Spain*, 2004.

[11] J. Kosecka, X. Yang, “Global localization and relative positioning based on scale-invariant features,” *In Proceedings of International Conference on Pattern Recognition*, 2004.

[12] J. Kosecka, X. Yang, “Location recognition and global localization based on scale invariant features,” *Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision*, 2004.

[13] A. Davison and D. Murray, “Simultaneous localization and map building using active vision,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 865-880, 2002.

[14] A. Davison, “Real-time simultaneous localization and mapping with a single camera,” *In Proceedings of the 9th International Conference on Computer Vision, Nice, France*, 2003.

[15] 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.

[16] L. Goncalves et al., “A visual front-end for simultaneous localization and mapping,” *2005 International Conference on Robotics and Automation, Barcelona, Spain*, 2005.

[17] N. Karlsson et al., “The vSLAM algorithm for robust localization an d mapping,” *2005 International Conference on Robotics and Automation, Barcelona, Spain*, 2005.

[18] D. Nister, “A minimal solution to the generalized 3 point pose problem,” *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2004)*, *Washington, DC*, *USA*, 2004.