AN INTEGRATED FEATURE SELECTION STRATEGY FOR MONOCULAR SLAM

[1] *Kuan-Ting Yu* (*俞冠廷*), [1] *Jia-Yuan Yu* (*余嘉淵*), [2] *Feng-Chi Liu* (*劉峰志*), [1] *Shih-Huan Tseng* (*曾士桓*), [1,2] *Li-Chen Fu* (*傅立成*), *and* [3] *Hsiang-Wen Hsieh* (*謝祥文*)

[1] Dept. of Computer Science and Information Engineering, National Taiwan University,
[2] Dept. of Electrical Engineering, National Taiwan University
[3] Mechanical and Systems Research Laboratories, Industrial Technology Research Institute
E-mail: r99922070@ntu.edu.tw

ABSTRACT

For feature-based monocular bearing-only SLAM, how to select useful features for SLAM process is crucial. The reason is overwhelming feature number will not only seriously slow down the system but produce erroneous SLAM result due to feature mismatching. In this paper, we propose a novel method for feature selection. The method combines both bottom-up (visual saliency) and top-down (learned object database) approaches to select versatile features. We argue that using human's visual saliency to guide the robot's visual SLAM feature selection is practicable. The experimental result after 10 runs attested our perspective. Compared with SLAM without feature selection, the running time here is reduced to 62% and the localization errors in the SLAM process decrease to 89% in mean and 89% in standard deviation.

Keywords Visual SLAM; Bearing only SLAM; Visual Saliency

1. INTRODUCTION

In order to perform human's orders in unstructured environments, it is crucial for robots to identify their current positions and construct the map of the surrounding areas. In the field of robotics, simultaneous localization and mapping (SLAM) research is regarded as the highly urgent issue to be solved. Due to the rich information carried by images and the low cost of a monocular camera, researchers are using monocular camera as a primary sensor for SLAM. The main objective of this paper is to address the following challenge of using monocular camera.

The typical way to construct the map is to extract robust feature points from input images as landmarks, which is called feature-based SLAM. To effectively select robust features with distinct, scale-invariant and viewpoint-adaptive property is crucial for the SLAM procedure, because processing time of typical SLAM algorithm scales with the number of features.

1.1. Related Work

Davison *et al.* [1] presented a MonoSLAM system which utilized Shi and Tomasi interest operator [3] to select features and solve. Civera *et al.* proposed to represent landmark in inverse depth. The inverse depth parametrization enables undelayed landmark initialization and possesses linearity property for EKF SLAM [4].

Lowe [5][6] presented SLAMB system using SIFT (Scale Invariant Feature Transform) [7] algorithm to generate 3D landmark points from a single robot pose with three cameras. Hundreds of matched points were used per image, with the database eventually storing many thousands of match points.

vSLAM®, a commercial product developed by Evolution Robotics company [8], combines both vision and odometry [9][10]. The system uses SIFT features, and this system is tested in a typical home environment and shown to produce reasonably accurate map.

Feature extraction is a key component that directly affects the ability of the system to reliably track and redetect features. Jensfelt *et al.* [11] used Harris-Laplacian detector [12]. Newman and Ho [13] used maximally stable extremal regions (MSER) [14]. Frintrop and Jensfelt [16] presented a biologically motivated feature extraction strategy to create a sparse set of feature. The idea is based on the principle of visual attention in the human visual system [17]. Their work inspires us to select fewer features from the regions of interest (ROIs) rather than the whole image. Thus, the SLAM process is more likely to achieve real-time performance. In this paper, we name it *bottom-up* selection strategy.

Fig. 1 Merge bottom-up ROI using human visual saliency (top-left) and top-down ROI using object detection (bottom-left) to select fewer yet useful features for monocular SLAM. The aim is to perform the SLAM system in an efficient and stable manner.

In practice, features selected from pure bottom-up saliency region are not always sufficient for monocular SLAM to achieve reasonable precision. In this situation, human can intuitively select some objects in the environment for the robot. Lyons utilized terrain spatiogram to combine RGB and spatial cues for landmarks using manually selected views [19]. Later, a landmark saliency architecture, LSA, includes visual, structural, and semantic attractiveness components to select candidate landmarks automatically [20]. Then, the robot can recognize those objects as features in the next SLAM execution. We name this procedure the *top-down* selection strategy.

Hochdorfe *et al*. presented a policy to rate and select landmarks in state vector based on landmark's coverage, which provides the information for evaluating the benefit of a landmark for localization [18].

Our proposed monocular SLAM system integrates both bottom-up and top-down features selection strategy to select fewer yet useful features, as shown in Fig. 1. Thus, we can perform SLAM more efficiently and still maintain stableness and acceptable precision for indoor application.

1.2. System Overview

Figure 2 shows the flowchart of the overall system. The inputs to the bearing-only EKF SLAM system are odometry and feature information obtained from the images. The outputs are the robot pose and a feature map. We propose the integrated feature selection algorithm to combine both bottom-up and top-down visual attention approach. The aim is to reduce the number of features to allow the EKF to perform efficiently and at the same time keeps the variety of the features.

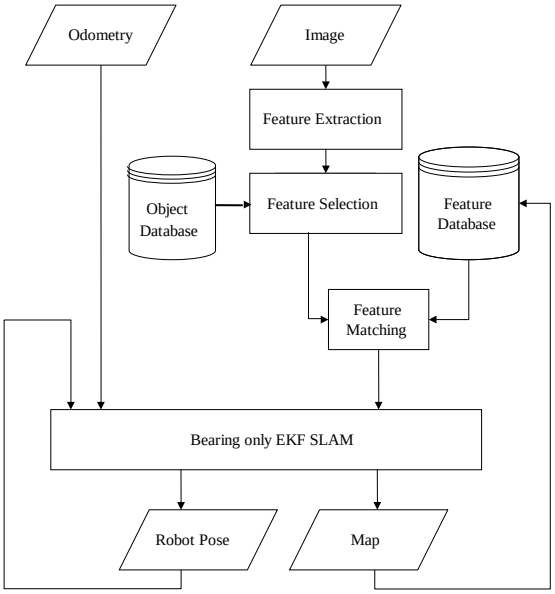


Fig. 2 Overview of the proposed monocular SLAM system with integrated feature selection strategy.

2. FEATURE EXTRACTION AND SELECTION

2.1. Feature Extraction

The Scale-Invariant Feature Transform (SIFT) [7] and Speeded Up Robust Features (SURF) [21] are two frequently used feature extraction algorithms for SLAM. SURF, with comparable repeatability and similar performance to SIFT [21], is a much faster interest point detector with one third of computation cost according to our experiments. Our system adopts SURF for feature extraction. However, the drawback is a huge number of features are extracted per image.

2.2. Feature Selection

An ideal candidate for selecting distinguishable regions in an image is a visual attention system [15]. The attention system selects features in ROIs which is similar to human visual system. It can considerably reduce the number of features stored in the feature database and the matching time.

There are two types of visual attention system, one is the bottom-up approach and the other is the top-down approach. Bottom-up approach uses the image-driven stimulus, while top-down is the knowledge-driven concept.

Many researchers are devoted to the study of bottom-up visual attention [16][22]. This saliency based regions selection strategy considers several feature channels independently, and strong contrasts or the uniqueness of features determine their overall saliency. The SLAM experimental result from Frintrop *et al.* [16] shows satisfying result using salient feature.

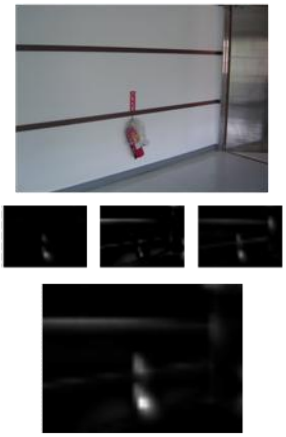


Fig. 3 A saliency image and its corresponding conspicuity maps

Fig. 4 The saliency image (left). The bottom-up ROI (right)

However, in practice, there are two problems for bottom-up saliency. 1) The bottom-up salient features are salient in image level but not in global level. For instance, the light bulb in an image is usually selected, but normally many identical light bulbs are in the room, which causes frequent mismatch in feature matching process. 2) Knowledge of object's movability is not known. Objects that are more possibly displaced, e.g., chairs, are considered unsuitable features in the map.

Hence, depending only on bottom-up visual attention is not enough. To build a feature map with global uniqueness and to prolong the valid time of the map, letting the robot also focus on objects that are static and globally unique, for example, sofa or pictures hanged on the wall, is beneficial. This is the top-down visual attention guidance.

2.2.1. Bottom-up Visual Attention

This algorithm consists of two typical processes. The first process is feature detection, where multiple low-level visual features, such as intensity, color, and orientation are extracted from the input image at different scales. The next process is saliency computation. The salient energy is obtained by a center-surround operation, which makes the salient region locally distinguishable. Thus, features from salient region tend to have lower mismatching rate. After normalization and linear combination of several conspicuity maps, salient image is generated to show which part is the most attractive. Note that the bright regions of the salient image are the regions of interest, which can be used for further analysis.

Fig. 5 Seven examples of the 2D images in object database. First row, from left to right: sign board, poster, fire-extinguisher and hydrant. Second row, from left to right: elevator, billboard and poster.

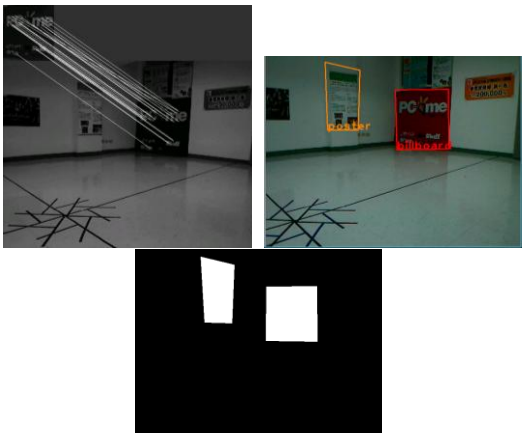


Fig. 6 The construction of top-down ROI: (top-left) feature matching for object detection, (top-right) two objects are detected in the scene, and (bottom) the final top-down ROI.

An implementation of the most popularly used bottom-up saliency model is proposed by Itti *et al.* [17]. There are some different modifications from the original model. The standard color space used as features in the original model is RGB color space. In our implementation, we choose the CIE L*a*b color space which was designed to properly mimic how human vision perceives the real world [24]. The saliency image generated from our modified model is shown in Fig. 3. We set a threshold to produce a binary mask of the last saliency image. The white regions are the ROIs, and the black area is ignored. One example is displayed in Fig. 4.

2.2.2. Top-down Visual Attention

Given an object database with 2D images as in Fig. 5, the robot can recognize them in a top-down manner during SLAM. The detection algorithm scans the input image captured from the camera and searches for known objects in its database. The detection process is based on SURF feature matching. First, FLANN (Fast Library for Approximate Nearest Neighbors) [25] matches the

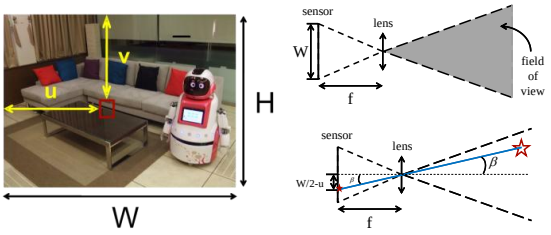


Fig. 7 Bearing information of selected feature.

features from the input image with the features from known object. Outliers in the matching results are eliminated with RANSAC (RANdom Sample Consensus) [26] method.

Through the pairs of corresponding points, we then solve the Homograph matrix [27] to get the projection function between these two images. Finally, we project the four end-points from the image patch to the current image to estimate the corresponding quadrilateral. The final top-down ROI is shown in Fig. 6.

2.2.3. Merge Two ROIs

Now we have two ROIs, one is from bottom-up visual attention and another is from top-down visual attention. We merge the two ROIs by OR operation to select more versatile SURF features, referring to Fig. 1.

2.2.4. Bearing Information of Selected Features

Using the pinhole camera model, we obtain the observed bearing β for the feature from the following transformation:

$$\beta = \arctan(\frac{(W/2)-u}{f}) \quad (1)$$

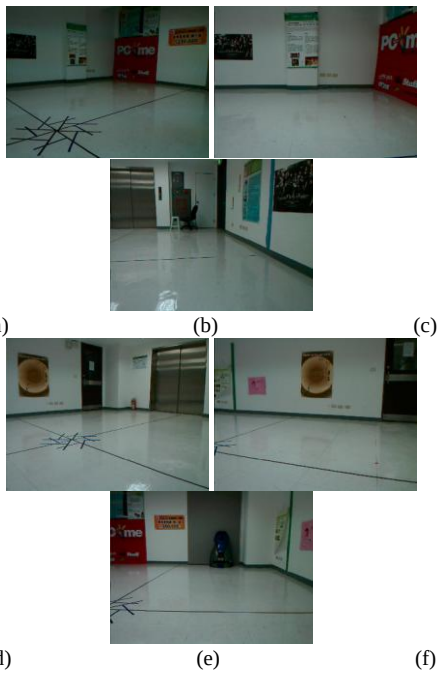
where (u,v) is the feature position on an image $W \times H$ and f is the focal length in pixel units, as shown in Fig. 7.

3. BEARING-ONLY SLAM WITH EKF

The prediction and update loop of the EKF per frame are as follows [28] :

1) *EKF Predict:*
   Perform the procedure of prediction according to the motion model.
2) *Landmark Extraction:*
   After the camera captures an image, SURF features are extracted and then selected based on our integrated visual attention system. Next, the bearing information is calculated as described in section 2.
3) *Feature Data Association*
   Chi-square test is used to determine the association between features from the current image and past images.
4) *EKF Update for initialized landmarks:*
   New measurements of existing map features are processed first in update.

Fig. 8. The experiment setting the monocular camera is mounted on the P3DX platform. The sick laser collects data for laser SLAM as ground truth.



(a) (b) (c)

(d) (e) (f)

Fig. 9 Some sample images of the SLAM database, with timestamp (a) 7, (b) 40 (c) 71, (d) 91, (e) 122, and (f) 152.

5) Check the SLAM state vector:

If there is existing a well-conditioned pair of measurements for a non-initialized landmark, the initial landmark estimate is added to the SLAM state vector [2]. Once a newly initialized feature is generated, the remaining stored measurements are applied for update. As soon as the observation poses do not produce unprocessed measurements, they are removed from the SLAM state vector. Finally, if the current robot pose provides a measurement not yet being initialized landmark, the observation is stored and the current robot pose is added as new observation pose to the SLAM state vector.

4. EXPERIMENTAL RESULTS

4.1. Visual SLAM Dataset

The experiments are performed on a P3DX platform. Logitech webcam V-UBH44 with horizontal view angle

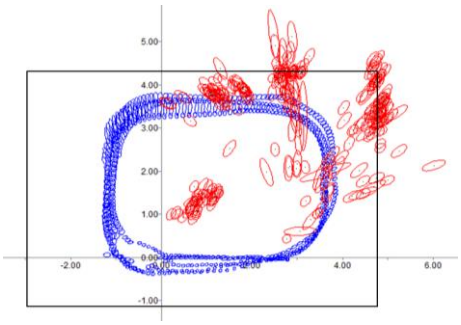


Fig. 10 Monocular SLAM with the proposed feature selection produced the feature map (red), and the estimated robot trajectory (blue). We marked the room boundary manually in black and ground truth trajectory with dashed line.

63 degree is mounted at the height of 80cm. The camera faces 45 degree to the front left, which aims to maximize both observing ability and parallax of the feature. A SICK laser range finder collects data to perform RBPF SLAM as ground truth.

We recorded data including images, odometry, and laser scanner reading for each time step. The robot was running at the speed around 110mm/s and captures data for every second. The experimental environment is a real-world indoor environment in Ming-Da Hall in National Taiwan University. The room size is about 6m × 8m. Fig. 9 shows some samples of our SLAM database. The robot circled a path of round corner rectangle for 4 times. After traveling a distance of 67.2 meters, the robot recorded 619 steps of data.

The monocular SLAM implementation is modified from Schlegel's bearing-only SLAM using omnidirectional vision [23]. The parameters of the motion model are $(0.03m)^2/1m$ noise for change in position and $(3deg)^2/360deg$ rotational error. The observation model uses $(0.5deg)^2$ as angular errors of the features. SURF extraction hessian threshold is set to 2000. The computation was run on a 2.4Ghz Intel Core2 Quad CPU.

The top-down ROIs are amended by hand after the automatic detection was done, producing nearly ideal top-down ROIs. We assume that a more sophisticated object detection system can allow this process completely automatic and with high performance.

4.2. Evaluation and Discussion of Feature Selection

We conducted the experiments using the following 4 feature selection strategies: 1) without selection (NO), 2) bottom-up (BU), 3) top-down (TD), and 4) the proposed integrated bottom-up and top-down selection (TDBU). We ran each selection method for 10 times. Fig. 10 shows an example of result of monocular SLAM using TDBU.

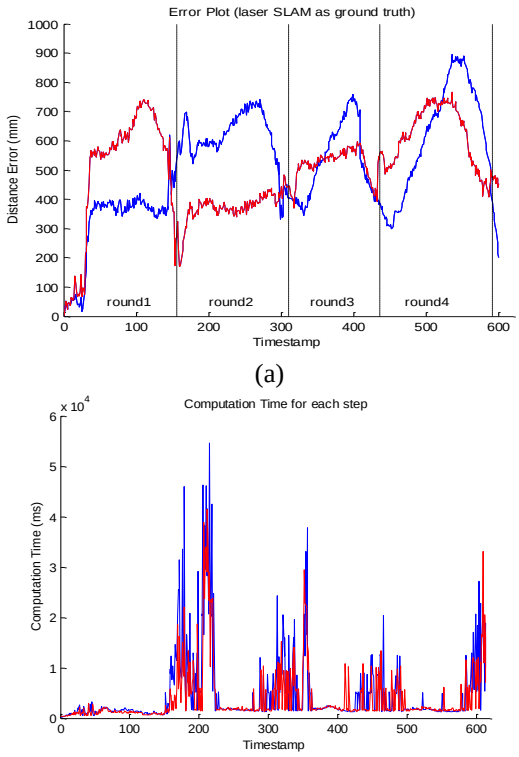


Fig. 11. (a) Localization error against ground truth. (b) Runtime of each step. The proposed TDBU masking is in red and NO masking is in blue.

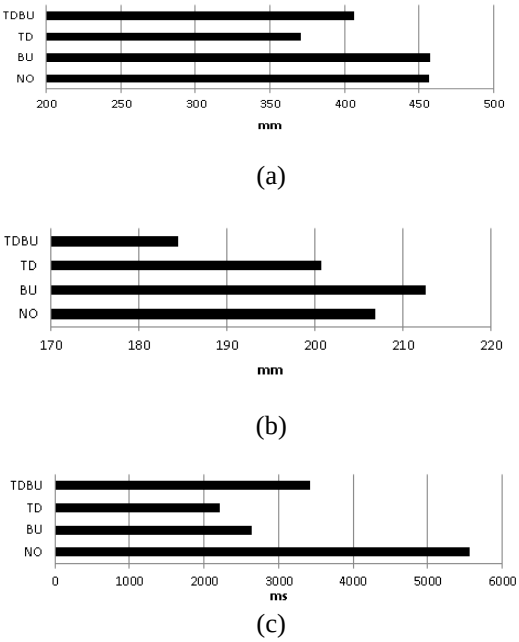
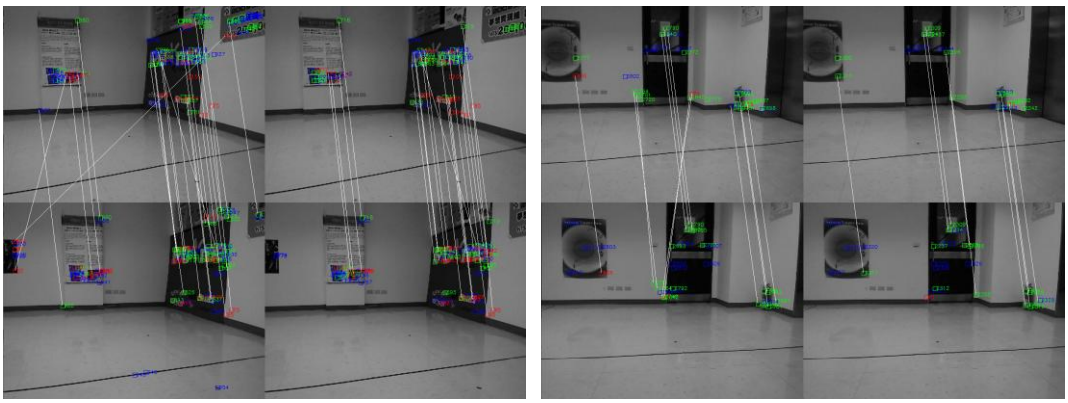


Fig 12. (a) Average translation error over one run. (b) Standard deviation of translation error over one run. (c) Average running time for each step. These comparison figures show average values after we run each method for 10 times.



(a) (b)

Fig. 13. Comparison of Monocular SLAM feature matching result of consecutive images at (a) timestamp 28 and (b) timestamp 100. The left part is the result without masking (NO) and the right part is the result with the proposed integrated feature matching (TDBU).

4.2.1. Accuracy Analysis

Figure 11 (a) shows the NO and TDBU translation error for one particular run. There is no obvious difference between NO and TDBU. Both have the similar pattern: the error is relatively small when the robot passed through the origin, at which more features with low uncertainty were initialized. Then error gradually increased after the robot left the origin.

To clearly compare the accuracy, we use the average translation error and standard deviation of translation error as the metric. Figure 12 (a) compares the mean error. BU has almost the same accuracy as NO, whereas TD has the best accuracy probably because the object we selected has the property of global uniqueness and right now the TD mask is ideal. The TDBU have the second accuracy performance. From the standard deviation

comparison in Fig. 12 (b), the TDBU has the smallest value, which means the SLAM error is relatively stable in the whole procedure. By examining the procedure, we find that in some views with BU or TD alone, too few features are selected and therefore cause the instability of SLAM.

4.2.2. Time Analysis

Figure 11 (b) shows the running time of NO and TDBU for one particular run. Figure 12 (c) shows the average running time. After feature selection, the SLAM computation time drastically decreases. TD and BU consumes respectively 40% and 47% time against NO, whereas the proposed TDBU uses 62%. In average, 57 features are extracted without selection, and 44 features are selected with the proposed selection strategy.

Note that the computation time here does not include the time for making BU or TD mask. One can easily include the computation time for a specific implementation for BU or TD.

4.2.3. Feature Matching Comparison

By visually inspecting the feature matching of the whole SLAM procedure, our feature selection strategy indeed abates the probability of feature's mismatching. Figure 13 shows two examples where incorrect feature matching in NO is avoided using TDBU feature selection.

5. CONCLUSION AND FUTURE WORK

We proposed a new integrated feature selection strategy for bearing-only SLAM with EKF by combining both bottom-up visual saliency and top-down recognition concept to obtain the versatile features. This visual attention system is applied to construct the ROIs to select fewer SURF features for monocular SLAM.

In the experiments, the number of selected SURF features is 77% of the original extracted number. We have shown that the proposed integrated feature selection strategy helps produce better localization accuracy and stability, and requires 62% SLAM running time compared with SLAM without feature selection. Therefore, we can conclude that using human's visual saliency to guide the robot's visual SLAM feature selection is practicable.

In the current trend of mapping application, robot should not only build a metric map with distance relations but create a semantic map with object information. For our current system, the features selected by top-down masking can be easily tagged with the object IDs and therefore achieve semantic mapping. Currently, appropriate top-down objects should be flat objects, e.g. signboards, posters, but the system is possible to be extended to 3D scenarios. To scale to large environment, some hierarchical map building techniques can be applied to fuse multiple local maps into a global map.

ACKNOWLEDGMENT

This work was supported by the National Science Council of the Republic of China Grant NSC 99-2221-E-002-190 and Taiwan Industrial Technology Research Institute FY99 Project: Visual Localization in Home Environment.

REFERENCES

[1] Andrew J. Davison, Ian D. Reid, Nicholas D. Molton, and Olivier Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, pp. 1052-1067, 2007.

[2] T. Bailey. Constrained initialisation for bearing-only SLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1966-1971 vol.2, Taipei, Taiwan, 2003.

[3] S. Jianbo and C. Tomasi. Good features to track. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 593-600, Seattle, WA, USA, 1994.

[4] J. Civera, A. J. Davison, and J.M.M. Montiel. Inverse depth parametrization for monocular SLAM. *IEEE Transactions on Robot.*, vol. 24, no. 5, Oct. 2008.

[5] S. Se, D. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2051-2058 vol.2, Seoul, Korea, 2001.

[6] S. Se, D. Lowe, and J. Little, Vision-based global localization and mapping for mobile robots. *IEEE Transactions on Robotics (T-RO)*, vol. 21, pp. 364-375, 2005.

[7] D. G. Lowe, Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, vol. 60, pp. 91-110, 2004.

[8] [webpage] Evolution Robotics: http://www.evolution.com/

[9] L. Goncavles, E. Di Bernardo, D. Benson, *et al*. A visual front-end for simultaneous localization and mapping. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 44-49, Barcelona, Spain, 2005.

[10] N. Karlsson, E. Di Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M. E. Munich. The vSLAM Algorithm for Robust Localization and Mapping. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 24-29, Barcelona, Spain, 2005.

[11] P. Jensfelt, D. Kragic, J. Folkesson and M. Bj¨orkman. A framework for vision based bearing only 3D SLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1944-1950, Orlando, Florida, USA, 2006.

[12] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 525-531, Vancouver, British Columbia, Canada, 2001.

[13] P. Newman, K. Ho. SLAM-Loop Closing with Visually Salient Features. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 635-642, Barcelona, Spain, 2005.

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 384-393, UK, 2002.

[15] S. Frintrop and P. Jensfelt and H. I. Christensen, Attentional Landmark Selection for Visual SLAM. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2582-2587, Beijing, China, 2006.

[16] S. Frintrop and P. Jensfelt, Attentional Landmarks and Active Gaze Control for Visual SLAM. *IEEE Transactions on Robotics (T-RO)*, vol. 24, pp. 1054-1065, 2008.

[17] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 20, pp. 1254-1259, 1998.

[18] S. Hochdorfer and C. Schlegel. 6 DoF SLAM using a ToF Camera: The challenge of a continuously growing number of landmark. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2009.

[19] Lyons, D.M., Sharing Landmark Information using Mixture of Gaussian Terrain Spatiograms. *IEEE/RSJ International Conference on Intelligent RObots and Systems (IROS)*, St Louis, MO, October 2009.

[20] Lyons, D. M., Selection and Recognition of Landmarks using Terrain Spatiograms. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2010.

[21] H. Bay, A. Ess, T. Tuytelaars, and Luc Van Gool. SURF: Speeded Up Robust Features, *Computer Vision and Image Understanding*, Vol. 110, No. 3, pp. 346--359, 2008.

[22] Y.-J. Lee and J.-B. Song, Visual SLAM in Indoor Environments Using Autonomous Detection and Registration of Objects. In *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2009.

[23] [webpage] Smart SLAM: http://smartslam.sourceforge.net/

[24] F. H. Imai, M. R. Rosen, and R. S. Berns, Comparative Study of Metrics for Spectral Match Quality. In *Proceedings of the European Conference on Colour Graphics, Imaging, and Vision*, France, 2002.

[25] M. Muja, and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International conference on computer vision and applications*, 2009.

[26] M. A. Fischler and R. C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, vol. 24, pp. 381-395, 1981.

[27] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[28] C. Schlegel and S. Hochdorfer. Localization and Mapping for Service Robots: Bearing-Only SLAM with an Omnicam. Advances in Service Robotics, pp. 253–278, 2008.