Accurate Visual Loop-Closure Detection Using Bag-of-Words for Multiple Robots

Jung H. Oh, Seung-Hwan Lee, and Beom H. Lee
Department of Electrical and Computer Engineering, Seoul National University, Seoul, Republic of Korea
Email: {bulley85, leeyiri1, bhlee}@snu.ac.kr

***Abstract*—We propose a method to detect loop-closures in a simultaneous localization and mapping (SLAM) problem for multiple robots. Each robot should be able to detect other robots' previously visited locations from camera measurements. To identify these places, our approach adapts the bag-of-words method in image recognition, and improves it by applying a Gaussian filter and a logistic function to correct the similarity scores. We can detect the robust loop-closures using only visual information of multiple robots. Experiments are performed to verify the effectiveness of the proposed method in indoor environments.**

***Index Terms*—loop-closure, mobile robots, SLAM, visual feature, bag-of-word**

I. Introduction

Humans and animals use visual information as a primary means of navigating the world and obtaining the environmental information. In robotics, visual information has been widely used as a solution to the simultaneous localization and mapping (SLAM) problem [1]–[3]. As cameras have become more compact and accurate while providing a rich qualitative description of the environment, they can replace the common range and bearing sensors such as laser scanners and sonars.

In SLAM, recognizing a place that has already been visited in a cyclical excursion or arbitrary length is referred to as loop-closure detection [4]-[6]. Such detection makes it possible to increase the precision of the pose estimate and global localization results by correcting the problem of error accumulation. Therefore, detecting the loop-closure not only improves SLAM performances, but also it enables additional capabilities to mobile robots.

The important point of the loop-closure detection problem is that it should not return any false positives. This is due to the fact that even a single false positive can cause irremediable failures during the SLAM procedure. Therefore, a good loop-closure detection method must not return any false positives while obtaining a minimum of false negatives.

In this paper, we extend the single robot visual loop-closure detection problem to multiple robots. Generally, using multiple robots can perform tasks more quickly and robustly than a single one [7]. Therefore, there have been approaches to solve SLAM problem with a team of cooperative robots as researched in [4], [8]-[9] and loop-closure in multi-robot mapping [10]. In particular, the loop-closure events, which are classified as rendezvous and map matching events, are analyzed at the global graph level in [10]. The concept of multiple-robot loop-closure detection is shown in Fig. 1.

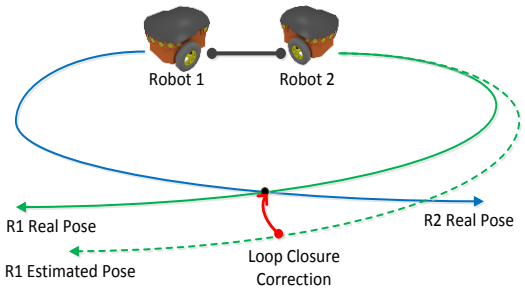


Figure 1. Loop-closure detection for multiple robots. Each robot can perform loop-closure correction from other robots' information

We used the bag-of-words method to perform loop-closure detection, as introduced in [4], [11]-[12]. We extracted features from images and construct a universal visual dictionary (UVD) by clustering a set of these unordered elementary features. Then, we were able to count the occurrence of the words and generate a histogram for each image. Comparing these histograms, we were able to calculate the similarity scores between images to find previously visited places of other robots'. To improve accuracy, we used the Gaussian filter and logistic function to correct the score and we were able to detect the robust loop-closure points using only visual information for multiple robots.

The remainder of the paper is structured as follows. We present a review of related work on visual SLAM and loop-closures for multiple robots in Section II. The method of generating UVD is presented in Section III and the details of an accurate loop-closure detection technique using the bags-of-words methodology is described in Section IV. To verify the proposed method, experimental results are given in Section V and finally, Section VI summarizes the conclusions.

II. Related Work

A summary of the SLAM problem is introduced in [1]. Previously, researchers were not focused on vision

Manuscript received July 10, 2014; revised October 17, 2014.

doi: 10.12720/joace.3.5.354-359

sensors, but used the laser range-finders or wheel encoder sensors. As the cameras have become more compact and accurate, intense research has been undertaken into visual SLAM employing cameras as primarily sensors. The term visual SLAM is implemented in [3] and [13]. The development of obtaining range information, environmental appearance, color and texture based on cameras give a robot the possibility of integrating other high-level tasks such as detection and recognition of people and places [14].

In the SLAM problem, detecting loop-closures has been one of the greatest issues to perform large scale SLAM and recovery from critical errors [4]-[6]. In [5] and [15], loop-closure detection is performed using an extended Kalman Filter (EKF) application to visual SLAM. According to [15], the method of detecting loop-closures is divided into three categories, map to map, image to image and image to map. In [4], the author proposed to use a similarity matrix to explain the relationships of similarities between all the possible pairs in captured images. In this paper, we use image to image similarity using the similarity matrix to find the loop-closures between the collections of image data from two robots.

To calculate the similarity between images, the bag-of-words method is applied in this paper, as introduced in [16-20]. This method is proposed to resolve data association in visual SLAM. In [16], the method of bag-of-visual words is proposed and [19] improved the method by applying the concept of the vocabulary tree. In computer vision community, the bag-of-word method is popular because it finds similar images at high speed. However, it also has the problem of detecting false positives and this is a very crucial problem in finding loop closures. To solve this problem, the epipolar constraint is used in [21] and conditional random fields (CRF) are used in [22]. Both methods use spatial information in the last phase of retrieval to find the loop-closures.

The bag-of-words method describes images as a set of unordered elementary features called visual words taken from a dictionary. There are different kinds of ways of representing images as features, such as scale invariant feature trans-form (SIFT) [23], speeded up robust features (SURF) [24], histograms of oriented gradients (HOG) [25], and so on. We used SIFT for the feature descriptors because of their robustness to reasonable 2-D affine transformations, scale, and viewpoint changes. The dictionary is built by clustering these similar visual descriptors extracted from the images into visual words. This is referred to as a universal visual dictionary (UVD). Using a given dictionary, we can count the occurrence of the words and generate a histogram for each image. Comparing these histograms, we can calculate the similarity scores between images. There are also other ways to construct UVD. In [21] and [26], vocabulary is constructed dynamically from features that are found as the environment is explored in real-time.

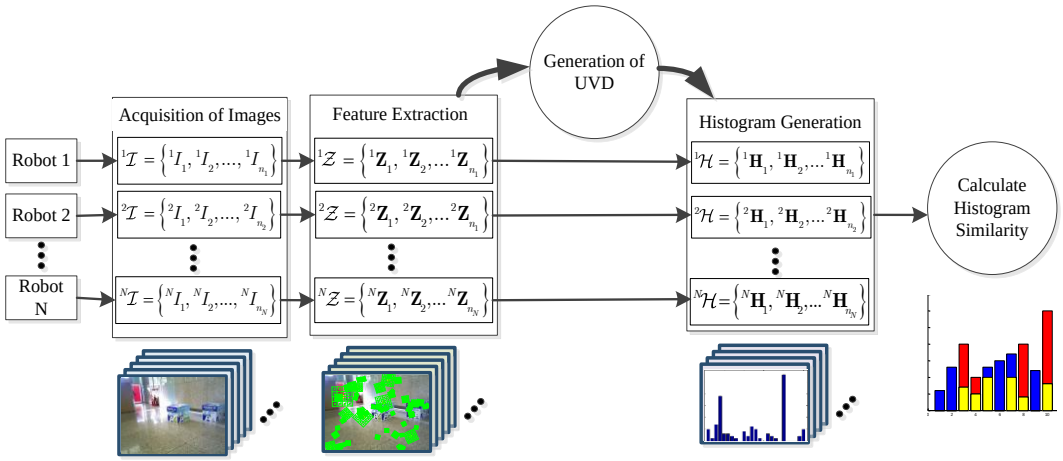


Figure 2. The overall procedure for generating the visual similarity matrix. Feature extraction is performed from acquired images of individual robot and the aggregated features are used to generate a universal visual dictionary. Each feature is quantized to the nearest visual words and represented by the histogram of visual words in universal visual dictionary. We calculate the similarity score using the histogram intersection method and generate the similarity matrix.

III. The Universal Dictionary Generation

The implementation of the bag-of-words method used in this paper is introduced in [17] and [18]. This method is based on vector quantization of affine invariant descriptors of images. We can construct the dictionary by clustering similar visual features. In this section, we first offer a method of extracting visual features from images and then, focus on generating UVD by clustering these features.

A. Extracting Visual Words

To find out the loop-closure points, we need to match the current image with a database of images acquired beforehand. Calculating the similarity as a whole image which is a set of pixels requires a large amount of memory and computation time. Therefore, we extracted features to reduce the dimension of inputs while containing the sufficient information to represent full size images. There are different kinds of feature descriptors

for representing images such as SIFT [23], SURF [24], HOG [25], and so on.

We used SIFT which is widely used as a descriptor of an image as it is consistent with 2-D affine transformations, variations of the illumination and other viewing conditions. Let ${}^{k}I_{t}$ be the obtained image at time *t* from robot *k* and ${}^{k}\mathbf{z}_{t}$ be the representation of this image in the feature space. Then each image can be associated with certain numbers of 128-dimensional feature vectors and we can deal with these vectors instead of raw images. We can denote this relationship as ${}^{k}I_{t} \Rightarrow {}^{k}\mathbf{z}_{t}$. If the features from images are not sufficiently large, we do not use these images to generate false positives errors in remaining procedure. As we obtained image queries from each robot, the set of images ${}^{k}\mathcal{I}$ which is obtained from robot *k* can be expressed as ${}^{k}\mathcal{Z}$ and denoted by ${}^{k}\mathcal{I} = \left\{ {}^{k}I_{1}, {}^{k}I_{2}, ..., {}^{k}I_{n_k} \right\} \Rightarrow {}^{k}\mathcal{Z} = \left\{ {}^{k}\mathbf{z}_{1}, {}^{k}\mathbf{z}_{2}, ..., {}^{k}\mathbf{z}_{n_k} \right\}$.

B. Generation of Universal Visual Dictionary

After extracting the features from images, we can generate the UVD by aggregating extracted features over all the images and clustering these features using K-means or KD-tree method. We used the K-means method which partition collected visual features into K clusters in which each feature belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Then, we can identify the size K of representative feature vectors which are called visual words and the set of these features are called universal visual dictionary. Using this dictionary we can compare the similarities between the obtained images for each robot.

IV. Loop-Closure Detection for Multiple Robots

The purpose of detecting loop-closures in the case of multiple robots is to recognize locations where other robots previously visited. In this section, we investigate ways of calculating the similarity between images based on the bag-of-words to identify these places. We also propose an enhanced model combined with the sequence probability information. Using this method, we are able to accurately detect the loop-closures based on visual information of two robots.

A. Bag-of-Words Representation

After generating UVD, each feature extracted from images can be quantized to the nearest visual words and can be represented by the histogram of visual words in UVD. We used the k-nearest neighbors (k-NN) algorithm to find the nearest visual words. Euclidean or Mahalanobis distance can be used for measuring the closest visual words in UVD.

Features are then extracted for an each image. Therefore, we can generate the histogram for each image, respectively. As the number of extracted features is different, histograms should be normalized to one. Let ${}^{k}H_{t}$ be the obtained histogram at time *t* from robot *k*. We can then represent each image as a histogram and a similarity score can be calculated from the similarity of these histograms.

B. Computing Similarity between Images

As we obtained the feature distribution of the image, we are able to classify the image from these histograms. To calculate the similarity between histograms, we used the histogram intersection function introduced in [20]. If we have the histograms H_i and H_j , the histogram intersection function $D(H_i, H_j) \in [0,1]$ can be represented as (1).

$$D(H_i, H_j) = \sum_{k} \min(H_i(k), H_j(k)) \tag{1}$$

If the two histograms are similar, the minimum value of each index will be large and the score will be high. On the contrary, if the two histograms are different, the score will be small. Therefore, we can use this measure as the criteria of similarity between histograms.

Now, we can create visual similarity matrix (VSM) between two robots using the histogram intersection scores. If we have images sequence ${}^{u}\mathcal{I}$ and ${}^{v}\mathcal{I}$ obtained from robot u and v , respectively, then the (i, j) of matrix VSM is expressed as (2).

$$S(i, j) = D({}^{u}H_i, {}^{v}H_j) \tag{2}$$

As $\left\| {}^{u}\mathcal{I} \right\| = N_u$ and $\left\| {}^{v}\mathcal{I} \right\| = N_v$, the size of VSM is N_u by N_v matrix. Using this matrix, we can code the relationships of resemblance between all the possible pairs in captured images. The example of the VSM is shown in Fig. 3. This is a similarity matrix constructed by comparing image sequences collected by two robots. We assumed two robots have traversed opposite directions. The dark line highlights the sequence of images that are similar to each other indicating that there is an overlap in the two environments explored. As the robots move in opposite directions, the diagonal entry shows a high similarity score.

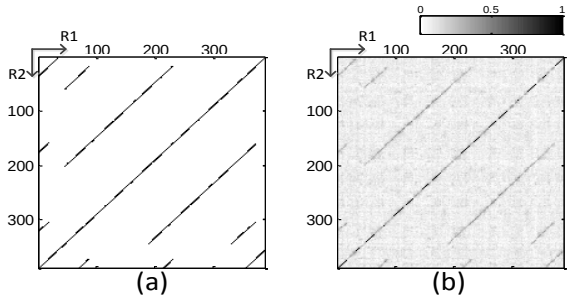


Figure 3. An example of generated VSM. Elements with high similarity scores are shaded dark while low similarity scores are not shaded. We assumed two robots have traversed through the same area in opposite directions in Lip6 dataset [21] (a) Ground truth (b) Generated VSM.

C. Accurate Loop-Closure Detection Method

In this section, we discuss how to detect loop-closures from this matrix by looking for sequences of similar images and the method of reducing false positives.

We obtained the similarity between all possible pairings of scenes from two robots in a VSM. We then pose the detecting loop-closures as the task of extracting statistically significant sequences of similar scenes from this matrix.

First, we apply the Gaussian filter to the VSM in order to reduce its sensitivity to noise. If the camera's field of view is narrow and robots move rapidly, the obtained image sequence might show a significant gap and features can be different even though it is a similar location. Therefore, the scores are not inconsistent in the same location. On the contrary, there can be many similar textures, such as repeated architectural elements, foliage and walls and some objects may appear in different places. This can cause false positives and make it difficult to recognize loop-closures. Therefore, we should observe neighborhood scores and this can be done using Gaussian filters. This effect is typically generated by convolving an image with a kernel of Gaussian values. The size of the filter and sigma value can be determined by the number of image sequences.

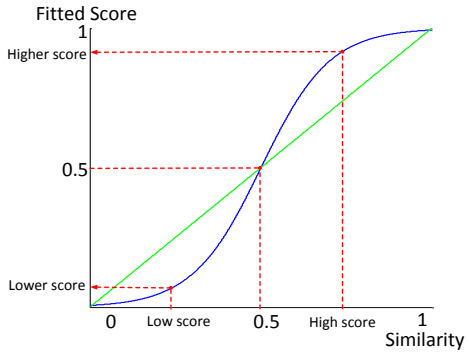


Figure 4. The logistic function in the case of $a = 10$, $c = 0.5$. As the score has a value between 0 and 1, we also have fitted values in the range of 0 and 1 by adjusting parameters.

Second, we fitted similarity score values using the logistic function. The function has the form of sigmoid function as shown in (3).

$$f(x) = \frac{1}{1 + e^{-a(x-c)}} \tag{3}$$

The function with two parameters a and c has the form of S-curve as depicted in Fig. 4. If x has a relatively small value, $f(x)$ will become smaller and if x has relatively large value, $f(x)$ will become larger value. The exact characteristics of this function are expressed in (4).

$$f(x) = \begin{cases} f(x) < x & if\ x < c \\ f(x) > x & if\ x > c \end{cases} \tag{4}$$

If we apply this function to the VSM, the low score gets lower value and the high score gets higher value. The reason for applying this step is to gain a higher contrast by changing the histograms of score values. This method spreads out the most frequent intensity values, therefore it allows for areas of lower local contrast to gain a higher contrast. As a result, it usually increases the global contrast and we can recognize the higher score areas more obviously. The results of the logistic function fitting are depicted in Fig. 5. We can see the distribution of the intensity gets flatter which results in increasing contrast of the scores.

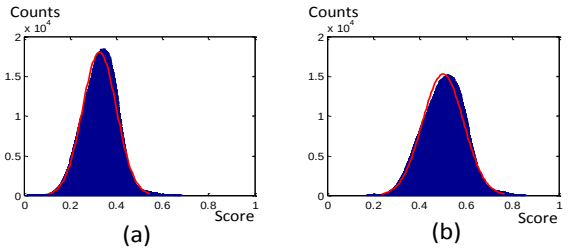


Figure 5. (a)The distribution of similarity scores before fitting. (b) The results of logistic function fitting. The distribution of the intensity gets flatter and the contrast of the image increases after fitting.

From these steps, we can extract significant sequence areas which have higher scores of similarity. Through this process, we were able to detect the robust loop-closure points using only visual information for multiple robots.

V. Experimental Results

To test and evaluate the proposed technique, experiments with two robots are conducted in an indoor environment. We assumed that illumination conditions remained constant during the experiment.

Each Pioneer-3DX robot is equipped with a RGB-D camera to collect the wheel encoder data and images. The total length of the experiment is 10 minutes and the frame rate is 10ms in 39×16 m ²area. The resolution of collected images is 640×480 and robots are traversed in opposite directions to detect the loop-closures as shown in Fig. 6. As we can see, the trajectory is shown with a dotted red line and the direction of each robot is shown as blue and green arrows. The robots are controlled using joystick trying to follow the same trajectory and obtaining data.

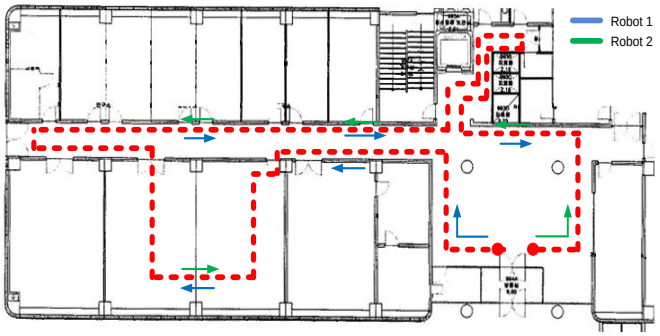


Figure 6. Overall robot trajectory for the image sequences. Two robots traversed in opposite directions to detect the loop-closures

Experimental results for the detection of loop-closures are shown in Fig. 7. The first image is a raw VSM obtained from SIFT and the bag-of-words method. Then, we filtered this image using Gaussian filter. The element of the matrix is changed according to the neighborhood pixels. As a result, the matrix is smoothed and noises are reduced. We apply a logistic-fitted function to this matrix as shown in Fig. 7 (c). The contrast of the matrix is increased which makes finding the loop-closure regions clearly. Finally, we can detect the loop-closures as depicted in Fig. 7 (d).

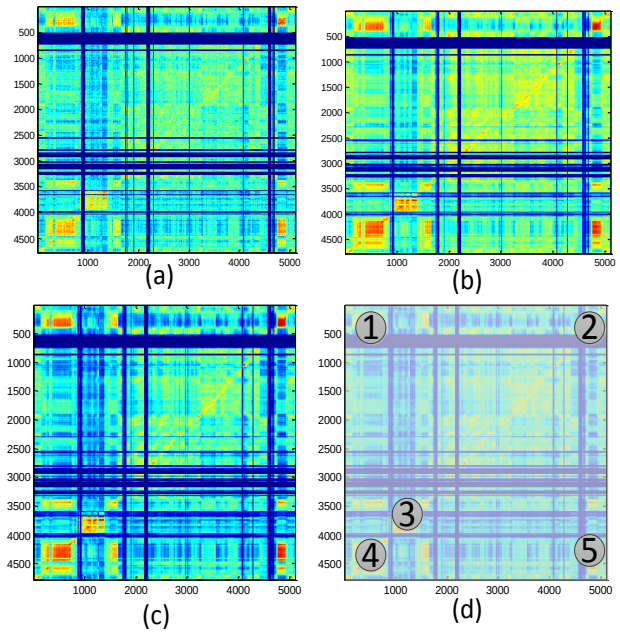
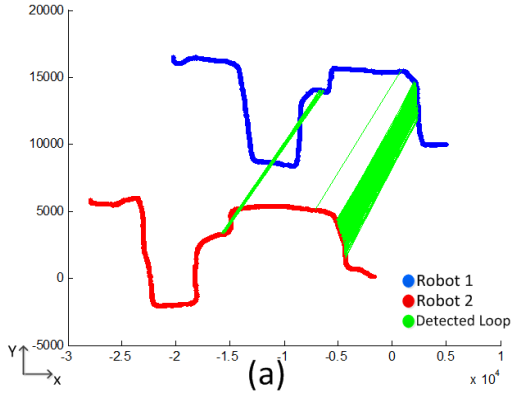


Figure 7. The experimental results for the detection of loop-closures using the bag-of-words method. (a) Raw VSM (b) Gaussian-filtered VSM. The matrix is smoothed and noises are reduced. (c) Logistic-fitted VSM. The contrast of the matrix is increased and we can find the loop-closure regions clearly. (d) Examples of detected loop-closures

From this matrix, we can find the loop-closure frames of each robot and connect the line as shown in Fig. 8 as we already obtained the trajectory of the robot from the wheel encoder. To present it clearly, we divided the trajectory into two parts and found the loop-closures respectively. The robots detect correct loop-closures while obtaining few false positives. Fig. 9 depicts the example images of detected loop-closures image from two robots. The obtained images of loop-closed connected illustrate similar images as expected.



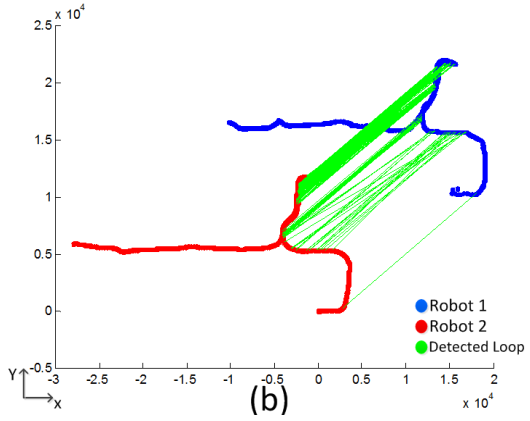


Figure 8. The trajectory of the robot and loop-closures results. The robots are controlled using a joystick to follow the same trajectory. (a) The half-way of the experiment. (b)The remaining part of the experiment.

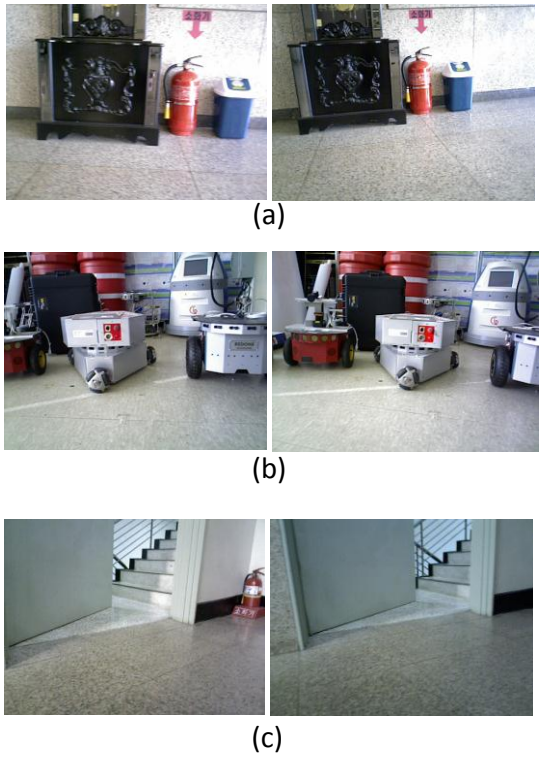


Figure 9. The detected loop-closures between two robots. (a) Robot 1 frame=4620, Robot 2 frame=212 (b) Robot 1 frame=1698, Robot 2 frame=3490 (c) Robot 1 frame=3442, Robot 2 frame=1578

VI. Conclusion

In this paper, we have presented a method for performing loop-closures in SLAM for multiple robots. To identify these places, we applied the bag-of-words method and extended it by integrating the Gaussian filter and logistic function to correct the similarity scores. To demonstrate the effectiveness of our approach, we conducted the experiments with two robots in indoor environments. The results did not return any false positive while a low number of false negatives were recorded. As a result, we were able to detect robust loop-

closures using only the visual information of multiple robots.

ACKNOWLEDGMENT

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIP) (No.2013R1A2A1A05005547), in part by the Brain Korea 21 Plus Project, in part by ASRI, in part by the Industrial Foundation Technology Development Program of MOTIE/KEIT [Development of CIRT (Collective Intelligence Robot Technologies)], and in part by Bio-Mimetic Robot Research Center funded by Defense Acquisition Program Administration [UD130070ID].

REFERENCES

[1] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robotics & Automation Magazine*, vol. 13, no. 2, pp. 99-103, June 2006.
[2] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, New York: The MIT Press, 2005.
[3] S. Se, D. G. Lowe, and J. J. Little, "Vision-based global localization and mapping for mobile robots," *IEEE Trans. on Robotics*, vol. 21, no. 3, pp. 364-375, June 2005.
[4] K. L. Ho and P. Newman, "Detecting loop closure with scene sequences," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 261-286, September 2007.
[5] L. A. Clemente, A. J. Davison, I. D. Reid, *et al.* "Mapping large loops with a single hand-held camera," *Robotics: Science and Systems*, vol. 2, 2007.
[6] C. Mei, G. Sibley, M. Cummins, *et al.* "RSLAM: A system for large-scale mapping in constant-time using stereo," *International Journal of Computer Vision*, vol. 94, no. 2, pp. 198-214, September 2011.
[7] W. Burgard, W, M. Moors, C. Stachniss, and F. E. Schneider, "Coordinated multi-robot exploration," *IEEE Trans. on Robotics*, vol. 21, no. 3, pp. 376-386, June 2005.
[8] A. Gil, Ó. Reinoso, M. Ballesta, and M. Julià, "Multi-robot visual SLAM using a Rao-Blackwellized particle filter," *Robotics and Autonomous Systems*, vol. 58, no. 1, pp. 68-80, January 2010.
[9] T. A. Vidal-Calleja, C. Berger, J. Solà, and S. Lacroix, "Large scale multiple robot visual mapping with heterogeneous landmarks in semi-structured terrain," *Robotics and Autonomous Systems*, vol. 59, no. 9, pp. 654-674, September 2011.
[10] T. A. Vidal-Calleja, C. Berger, and S. Lacroix, "Event-driven loop closure in multi-robot mapping," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2009, pp. 1535-1540.
[11] J. Wang, H. Zha, and R. Cipolla, "Coarse-to-fine vision-based localization by indexing scale-Invariant features," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 36, no. 2, pp. 413-422, April 2006.
[12] P. Newman, D. Cole, and K. Ho, "Outdoor SLAM using visual appearance and laser ranging," in *Proc. IEEE International Conference on Robotics and Automation*, May 2006, pp. 1180-1187.
[13] T. Lemaire, C. Berger, I. Jung, and S. Lacroix, "Vision-based SLAM: Stereo and monocular approaches," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 343-364, September 2007.
[14] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: A survey," *Artificial Intelligence Review*, Netherlands: Springer, November 2012.
[15] B. Williams, P. Smith, and I. Reid, "Automatic relocalisation for a single-camera simultaneous localisation and mapping system," in *Proc. IEEE International Conference on Robotics and Automation*, April 2007, pp. 2784-2790.
[16] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. IEEE International Conference on Computer Vision*, vol. 2, October 2003, pp. 1470-1477.
[17] G. Csurka, C. R. Dance, L. Fan, *et al.* "Visual categorization with bags of keypoints," in *Proc. European Conference on Computer Vision (ECCV)*, vol. 1, May 2004, pp. 1-22.
[18] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. IEEE International Conference on Computer Vision*, vol. 2, October 2005, pp. 1800-1807.
[19] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, June 2006, pp. 2161-2168.
[20] S. Lazebnik, S, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, June 2006, pp. 2169-2178.
[21] A. Angeli, S. Doncieux, J. A. Meyer, and D. Filliat, "Real-time visual loop-closure detection," in *Proc. IEEE International Conference on Robotics and Automation*, May 2008, pp. 1842-1847.
[22] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Proc. European Conference on Computer Vision*, 2010.
[23] D. G. Lowe, "Distinctive image feature from scale-invariant keypoint," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, November 2004.
[24] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346-359, June 2008.
[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, June 2005, pp. 886-893.
[26] T. Botterill, Tom, M. Steven, and R. Green, "Bag-of-words-driven, single-camera simultaneous localization and mapping," *Journal of Field Robotics*, vol. 28, no. 2, pp. 204-226, March/April 2011

Jung H. Oh received the B.S. and M.S. degrees in Electrical Engineering and Computer Science from Seoul National University, Seoul, Korea in 2012 and 2014. He is currently a Ph. D. candidate in the Department of Electrical and Computer Engineering at Seoul National University. His research interests include vision-based robotics applications, simultaneously localization and mapping, and multi-agent system coordination.

Seung-Hwan Lee received the B.S. degree in Electronic and Electrical Engineering from Kyung-Book National University in 2008. And he received the M.S. degree in Electrical Engineering and Computer Sciences from Seoul National University in 2010. He is currently a Ph.D. candidate in the Department of Electrical and Computer Engineering at Seoul National University. His research interests include SLAM;

Beom H. Lee received his B.S. and M.S. degrees in Electronics Engineering from Seoul National University, Seoul, Korea in 1978 and 1980, respectively, and his Ph.D. degree in Computer, Information and Control Engineering from the University of Michigan, Ann Arbor, in 1985. From 1985 to 1987, he was with the School of Electrical Engineering at Purdue University, West Lafayette, IN, as an Assistant Professor. He joined Seoul National University in 1987, where he is currently a Professor at the School of Electrical Engineering and Computer Sciences. Since 2004, he has been a Fellow of the Robotics and Automation Society. His research interests include multi-agent system coordination, control, and application.