https://doi.org/10.20965/jrm.2017.p0685

Paper:

Robust and Accurate Monocular Vision-Based Localization in Outdoor Environments of Real-World Robot Challenge

**Adi Sujiwo[*], Eijiro Takeuchi[*], Luis Yoichi Morales[**], Naoki Akai[**],
Hatem Darweesh[*], Yoshiki Ninomiya[**], and Masato Edahiro[*]**

[*]Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
E-mail: {sujiwo, eda}@ertl.jp, {takeuichi@coi, hatem.darweesh@g.sp.m.is}.nagoya-u.ac.jp
[**]Institute of Innovation for Future Society, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
E-mail: {morales_yoichi, akai, ninomiya}@coi.nagoya-u.ac.jp

[Received March 3, 2017; accepted June 6, 2017]

This paper describes our approach to perform robust monocular camera metric localization in the dynamic environments of Tsukuba Challenge 2016. We address two issues related to vision-based navigation. First, we improved the coverage by building a custom vocabulary out of the scene and improving upon place recognition routine which is key for global localization. Second, we established possibility of lifelong localization by using previous year's map. Experimental results show that localization coverage was higher than 90% for six different data sets taken in different years, while localization average errors were under 0.2 m. Finally, the average of coverage for data sets tested with maps taken in different years was of 75%.

Keywords: visual localization, field robotics, Tsukuba Challenge

1. Introduction

Reliable autonomous navigation in dynamic environments is a core competency for robot but complicated task given the different conditions that a robot has to face. Current state of the art navigation approaches usually rely on accurate vehicle localization to compute a path between their current position towards a destination. Localization in outdoor environments is a hard task given the highly dynamic conditions. It is specially difficult for visual systems to cope with the different environmental changes [1] given by light conditions, populated environments and change of place visual characteristics caused by the different seasons of the year [2–4].

The Real-World Robot Challenge (RWRC) is an annual robotic autonomous navigation challenge held in the public space of Tsukuba, Japan. The robots are required to navigate over a 1 km route, so maintaining localization accuracy within dynamic environment is necessary. Previous Tsukuba Challenge events had seen some applications of vision-based localization system (e.g., [5]). However, there have not been any teams who achieved the autonomous navigation task solely using vision. Almost all teams who achieved the navigation task used Light Detection and Ranging (LIDAR) for localization (e.g., [6]). These papers show that vision-based localization and navigation is still a challenging task. Meanwhile, our objective is to realize practical outdoor navigation using consumer-level camera. However, maps used during the navigation step are provided from a robot with high end sensors to build high quality maps and provide accurate 6 DoF poses. In other words, the focus of this study is how to achieve accurate localization using consumer level sensors with maps which include rich information. **Fig. 1** shows the main blocks of our proposal.

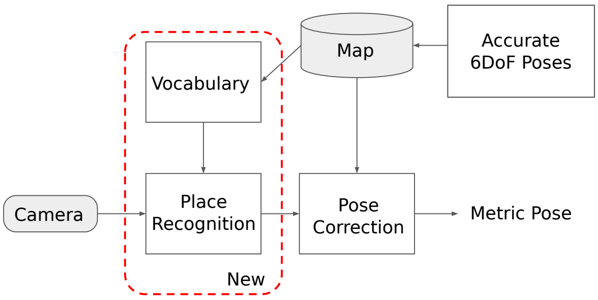


Fig. 1. Simplified block diagram to compute metric pose with a single monocular camera.

We have proposed a localization method using monocular camera in [7]. In previous paper, we modified mapping process of ORB-SLAM [8] and achieved monocular camera-based localization in accurate metric coordinate. The method allows frame-based fusion of monocular camera-based localization results and external metric-based sensors (e.g., GPS and odometry) by using particle filtering algorithm. We tested the method in Tsukuba Challenge 2015. Although the localization method was able to accurately estimate own position in some areas of Tsukuba-city, coverage of the method was not enough (localization coverage was approximately 67%).

The main differences of this work towards previous work in [7] are an improved localization coverage based on custom vocabulary and improved global localization routine. These two blocks are illustrated inside dotted lines in **Fig. 1**. We use preprocessing for building an exact visual feature map and modified visual vocabulary which is used for place recognition. These modifications improve localization accuracy and coverage of our visual localization method. Experiments using log data taken at Tsukuba Challenge 2015 and 2016 are used to demonstrate effectiveness of the proposal. In this study, performance of the proposed method is evaluated on the basis of the 3D LIDAR-based localization method that gives us exact estimation results which can be assumed as ground truth [9]. A simplified re-localization process which does not remove similar candidates improves global localization.

The contributions of this work are twofold:

- Experimental proof of coverage improvement from custom vocabulary of the outdoor scene.
- Enhanced re-localization routine which is key for global localization.

The rest of this paper is organized as follows. Section 2 summarizes related works. Section 3 and 4 present the monocular visual-based system for mapping and localization, and the design and implementation. The experimental procedure and results are described in Section 5 and 6. Section 7 concludes this work.

2. Related Works

There are some works which address autonomous navigation in pedestrian paths [10, 11] which show that long range navigation is feasible. Yet, outdoor navigation is a complicated task to achieve. Robot localization modules usually rely on environmental maps previously built. As environments change with time, map maintenance is necessary. Map update is a hard task given that consistent map building with the same coordinate frame is necessary. There are also existing works regarding visual map maintenance and update (e.g., [12]).

An interesting study of lifelong vision-based localization can be found in [13]. They referred a summary map that is built from several localization trials. This means that they tried localization experiments in the same place and updated a visual feature map. By the summary map, they succeeded in lifelong visual localization over 16 months. We proposed similar idea which uses multiple visual maps to cope with a problem of appearance change in previous work [7]. We kept consistency of the multiple visual maps by utilizing 3D LIDAR-based localization results.

Most of current method in robotic motion planning depends on accurate geometry of vehicle and its environment [14]. In this regard, motion planner algorithm usually search for most optimum paths that are subject to presence of obstacles and vehicle motion constraints. Due to this nature, planner algorithm requires that both localization and obstacle detector work in metric space. However, as stated in [15], current vision SLAM methods (and thus localization) are not free from scale drift due to their inherent limitations [16]. An example of solution of combined vision-based navigation that works in topological space is devised in [17] and [18].

In our previous paper [7], we have developed metrically accurate localization system that provides reasonable accuracy in metric space. This method works by augmenting keyframe positions in the vision map with the accurate metric pose estimated by the other 3D geometric map-based localization. Assuming that scale drift is small in vicinity, position estimation in metric space can be obtained by scaling the translation from keyframe.

A similar method to ours can be found in [19]. In this work, 3D reconstructed feature points from local bundle adjustment are matched with 3D LIDAR maps. Using accurate geometric maps to cope with the scale drift problem in monocular vision-based system is same idea. The method proposed here is a geometric-based matching method because 3D reconstructed features are directly matched with the LIDAR maps in metric frame. In contrast, our proposal is an appearance-based matching method because ORB-SLAM estimates own pose by comparing ORB features. These methods have different advantages from each other.

Current progress of visual place recognition for SLAM purposes have been surveyed in [20]. As mentioned in the paper, handling variable illumination conditions is critical for place recognition performance. One possible solution is by tweaking color-to-grayscale conversion; as shown in [21], this commonly ignored process has significant contribution for accuracy of image recognition. An interesting possibility instead of simple color conversion is to perform change image colors to illumination-invariant color space [22]; which can be used to tackle shadows and light changes throughout times of day.

3. Monocular Vision-Based Mapping and Localization

Most vision-based SLAM methods are based on Structure-from-Motion (SfM), as surveyed in [23] and [24]. General workflow of the SfM as implemented in ORB-SLAM is described in **Fig. 2**. Explanation of each major parts will be accompanied with details of their implementation in ORB-SLAM that we use.

This workflow is applied for mapping process. For localization process, generally the system does not include map modification so that local mapping and loop closure are omitted from the workflow.

3.1. Feature Tracking

First step in pose and structure recovery is detection of distinct feature points in the frame. In ORB-SLAM,

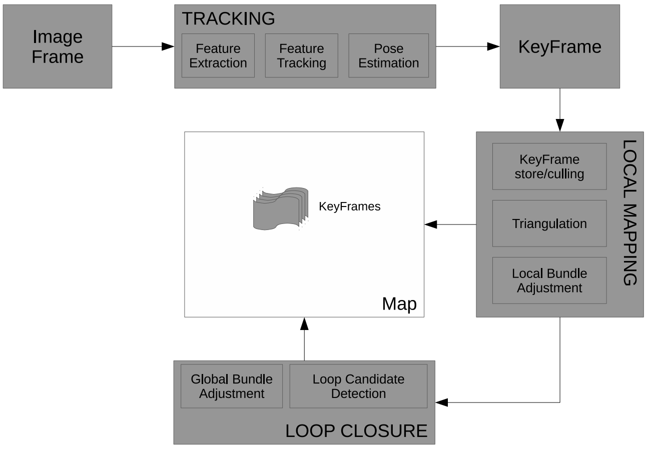


Fig. 2. General vision-based SLAM workflow.

ORB (Oriented FAST, Rotated BRIEF) is used as feature detector [25]. It has been found to be invariant against rotation, but not against scaling and illumination change [26]. Next, the system must match and track these feature points in subsequent frames, so that these points may be used for triangulation.

In our observation, the matching and tracking subsystem will provide better accuracy for high parallax features; this requires that tracked features lie in near places (e.g., low-height vegetations, paving tracks and signs). However, features in low parallax (i.e., far places) such as trees and buildings are also required for place recognition. Meanwhile, the feature detection and tracking must be prevented from matching features in bright skies and clouds, which may confuse visual odometry.

In practice, the camera is not always able to cope with these two competing requirements due to limited camera's dynamic range. This situation is prevalent in high contrast areas due to strong illumination in sky but dark shadows in the ground are looming. Therefore, in this research we insert an image preprocessing step that perform gamma correction based on histogram measurement of adjustable image portion (see **Figs. 3** and **4**).

3.2. Local Mapping and Triangulation

A number of distinct frames are recognized as keyframes when they contains enough changes from surrounding frames. Position and orientation of these keyframes are computed from fundamental matrix as described in [16]. Next, map points are computed from two consecutive keyframes using triangulation, also in [16]. The triangulation process has higher uncertainty in translation motion than rotation, which in turn makes corresponding keyframes and map points more difficult to be recognized. This uncertainty is shown in our previous paper, which describes no coverage in starting point that involved long and straight motion.

Due to inherent nature of 3D poses and points reconstruction that can be only computed up to an unknown scale [27, 28], the map will contains scale drift [15]; therefore it is necessary to correct this deficiencies. In the ORB-SLAM, scale drift is corrected by using local bundle adjustment, in which nearby keyframes and map points are adjusted to give cumulative projection error. However in our observation, the local adjustment may fail when most map points are present in objects that have low parallax (for example, cloud in the sky and/or faraway buildings). This situation comes for example, when high contrast between sky and ground are present in the scene.

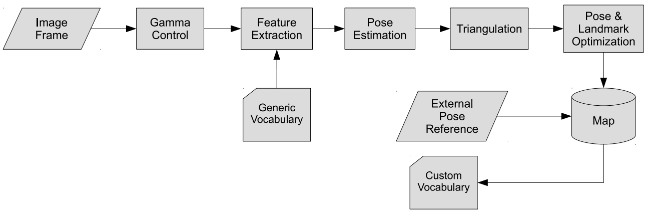


Fig. 3. Flowchart of the proposed mapping method. Construction of custom vocabulary from the final visual map is new addition from the previous method. We also inserted gamma control as image preprocessing prior to feature extraction.

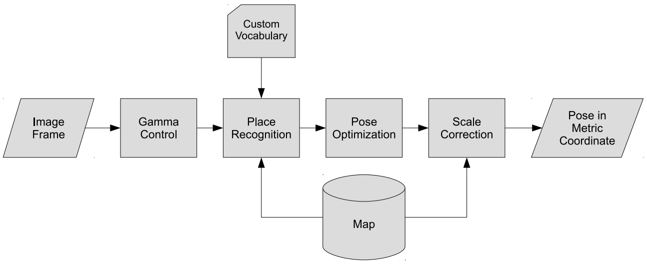


Fig. 4. Flowchart of the proposed localization method. Place recognition process uses custom vocabulary instead generic vocabulary. Similar to mapping, localization also uses gamma control for frame preprocessing.

3.3. Loop Closure and Place Recognition

Loop closure is crucial for enhancing accuracy of SLAM algorithms, determining whether the robot has returned to visited locations after discovering new terrain and proceeding to adjust keyframes and map points. In the ORB-SLAM, loop closure uses image-to-map approach [29]. This is performed by querying extracted ORB features of current keyframes against the database of bag-of-words [30], and performing global bundle adjustment against the whole map.

One critical element in loop closure is place recognizer that are also used in initializing localization process [20]. The place recognition enables ORB-SLAM to perform global localization, thus eliminating requirement of starting localization in point zero of the map.

3.4. Scale Correction to Metric Space

Our previous paper [7] has derived scale correction to transform poses in ORB-SLAM coordinate frame to metric space, by assuming that scale changes in small movements are uniform. This is performed by storing

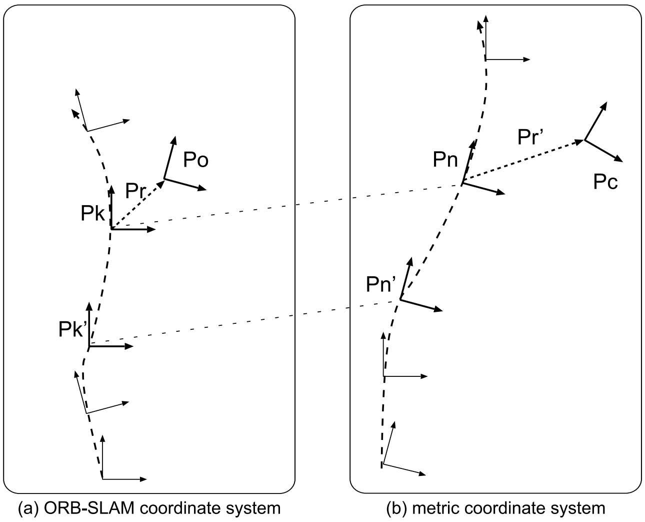


Fig. 5. Scale correction to metric space from ORB-SLAM frame.

keyframes' accurate metric pose estimated from an external localization method (e.g., using LIDAR) in the mapping phase.

Figure 5 shows illustration of correction to metric space. During mapping phase, keyframes' real poses in metric space are recorded along with their computed positions in ORB-SLAM frame as P_n and P_k, respectively. Each pose P consists of translation vector $\mathbf{t} = (x,y,z)^T$ and rotation matrix R. In localization phase, robot pose in metric frame as P_c is predicted as scaling-up from ORB-SLAM pose P_o. First, we search for nearest keyframe in ORB-SLAM frame as P_k and its offset as P'_k; their positions in metric frame are P_n and P'_n. Next, scale factor s is calculated as

$$s = \frac{||\mathbf{t}'_n - \mathbf{t}_n||}{||\mathbf{t}'_k - \mathbf{t}_k||}. \quad \ldots\ldots\ldots\ldots\ (1)$$

P'_r is transformation from P_n to P_c, computed by the following expression:

$$P'_r = \begin{pmatrix} R(\mathbf{q}_r) & s\mathbf{t}_r \\ \mathbf{0}^T & 1 \end{pmatrix}, \quad \ldots\ldots\ldots\ (2)$$

where $R(\mathbf{q}_r)$ is the rotation component of P_r. Lastly, final pose in metric frame is derived from:

$$P_c = P_n P'_r. \quad \ldots\ldots\ldots\ldots\ (3)$$

4. Proposed Improvements to Previous Localization System

Our previous paper has proposed metric-based vision based localization system. The readers are encouraged to refer to this paper for explanation of our modification to original ORB-SLAM and how to obtain pose estimation in metric space. In this paper, we propose additional improvements that aims to increase coverage. Main features of current addition are:

1) custom vocabulary for place recognition;

2) automatic gamma control; and

3) non-strict keyframe selection.

Framework of our localization system is shown in **Figs. 3** and **4**. These figures describe the map building and localization processes.

4.1. Use of Custom Vocabulary

Original ORB-SLAM employs a generic vocabulary extracted from an unspecified training image sequences [8], which was noted to work well for a number of publicly available datasets. As described in [30], this vocabulary is used for transforming detected features of the image into a sparse numerical vector (thus the name "bag-of-words"). In 2015 and 2016 Tsukuba Challenge, we found that it was not the case, as the place recognition may fail when using generic vocabulary. Therefore, the first proposed addition for ORB-SLAM is to utilize custom vocabulary for any specific location (in this one, the Tsukuba Challenge track) to increase probability of matching query image against the image database [31]. The vocabulary is extracted following the mapping process, as the image sequence is required.

Constructing image vocabulary is basically a form of vector quantization [32, 33], in which the vocabulary is arranged as tree. The process of extracting vocabulary is performed by collecting a rich set of feature descriptors from training images. As described in [30], the extracted descriptors are discretized and clustered using k-means and inversely weighted according its relevance in the training sequence. The whole vocabulary construction are processed by DBoW2 library [30].

4.2. Automatic Gamma Control

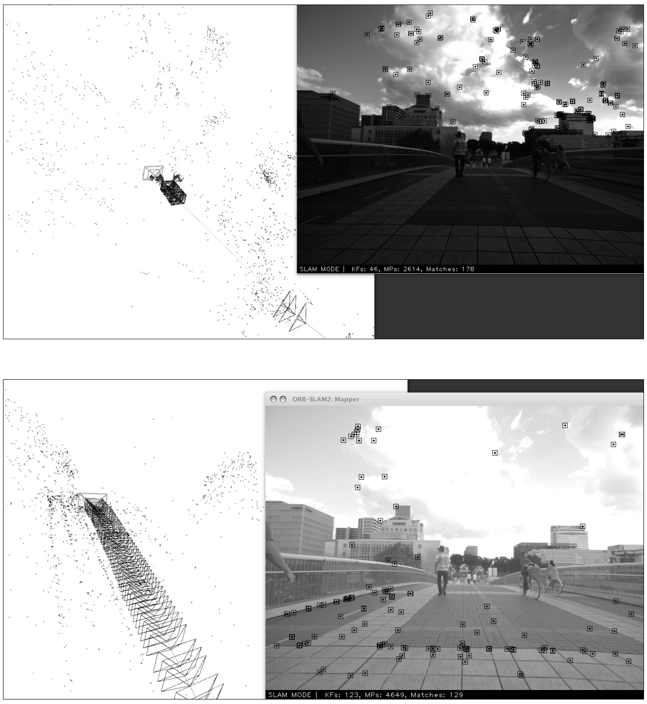
As described in [26], there are no feature detectors and descriptors that are illumination-invariant; hence feature matching may not work under varying brightness. We apply gamma correction to handle high contrast situation in daytime lighting as often encountered in the Tsukuba Challenge track. The gamma correction basically works by applying exponential correction for pixel value: $I_i \leftarrow I_i^\gamma$, where I_i $(0 \leq I \leq 255)$ is a pixel value of i-th pixel.

To automatically decide the value of γ, we first compute histogram of pixel values, $h(I)$, included in masked region on the image, A, denoted as:

$$h(I) = \sum_{i \in A} \delta_{I,I_i}, \quad \ldots\ldots\ldots\ldots\ (4)$$

where δ is Kronecker delta. Cumulative distribution function (CDF), $c(I)$, is then calculated from the histogram as:

$$c(I) = \frac{\sum_0^I h(I)}{\sum_0^{255} h(I)}. \quad \ldots\ldots\ldots\ldots\ (5)$$



SLAM MODE | KFs: 46, MPs: 2614, Matches: 178

ORB-SLAM2: Mapper

SLAM MODE | KFs: 123, MPs: 4649, Matches: 129

Fig. 6. Mapping without (top) and with (bottom) gamma correction. In top figure, almost all tracked ORB features fall in the skies and clouds, resulting in closely spaced keyframes but sparse map points; signifying relatively little motion (pyramid markers depict keyframes). In contrast; using gamma correction, result keyframes are uniformly spaced, and more map points fall in the ground with distinctive patterns following their placement in the ground.

Then we compute the value of γ to adjust for the midtone that aims to simulate human visual response against strong backlight [34] as:

$$\gamma = -\frac{\ln I_{50}}{\ln 2}, \quad \dots\dots\dots\dots\dots \quad (6)$$

$$I_{50} = c^{-1}(0.5). \quad \dots\dots\dots\dots\dots \quad (7)$$

The γ value is calculated from masked region which represents the midtone intensity of that region; however, we apply the gamma correction to whole image. The masked region may be determined arbitrarily; but the best results are obtained when it is taken from lower half of image, as this region is subject to be dark when the camera is facing high contrast scenes.

Effect of this gamma correction is to add brightness and contrasts in shadow areas, while reducing contrast in highlighted ones. In turn, there are more ORB features to detect and tracked in the ground (closer to camera). This is shown in **Fig. 6**.

4.3. Non-Strict Prediction for Relocalization

Original ORB-SLAM implementation stipulates relocalization by bag-of-words (BoW) search in internal database for looking up keyframe candidates. This set of candidates are then filtered by discarding similar keyframes and its preference to keyframes that have history of previous match with prior queries. The candidate filtering acts to reduce computation time, as the next step (scoring and geometry check) is quite expensive. In practice, this method often fails because either the number of candidates is too few, or the candidates do not match with geometry check. To increase success probability of relocalization, we propose modification of candidate selection by removing the candidate filtering. Instead, we compute scores of all candidates and select 25% best keyframes. Obviously, this necessitates trade-off between CPU usage and coverage.

To accelerate position finding, we also add searching nearest keyframes that share visible map points with last good keyframes. Consequently, this method is not usable for initializing global localization when the system starts, as no prior information of keyframes exists. The idea of searching nearest keyframes is not new; it is actually inspired by PTAM [35]. Here, we combine this method with BoW search as fallback. Complete algorithm for relocalization is listed in **Algorithm 1**.

Algorithm 1: Relocalization after Lost Occurrence

```
Data: Image Frame with descriptors
Result: Keyframe Candidate $c$ or $\emptyset$
if prior keyframe is not found then
    Find keyframes from database that share descriptors
    with image frame as $K$
else
    Search keyframes that have topological relation
    with last good keyframe as $K$
end
Compute scores for all in $K$
Select candidate $c$ with max. score
Geometry check:
    Project all keypoints in $c$
    if Keypoints in c match with image frame then
        return $c$
    else
        return $\emptyset$
    end
```

5. Experimental Procedure

We conducted three experiments to examine efficacy of our method. First experiment is used to validate performance characteristics in terms of accuracy and coverage against recent (2016) Tsukuba Challenge track. We also want to identify which situations may cause failure of our method. This information will be important for further deployment of vision-based localization in public road cases. The trajectory of 2016 Tsukuba Challenge is shown in **Fig. 7**. One mapping run and four localization runs were conducted separately for this experiment; all runs were from same timeframe. We also took ground truth measurements in both mapping and localization runs using results from LIDAR-based localization.

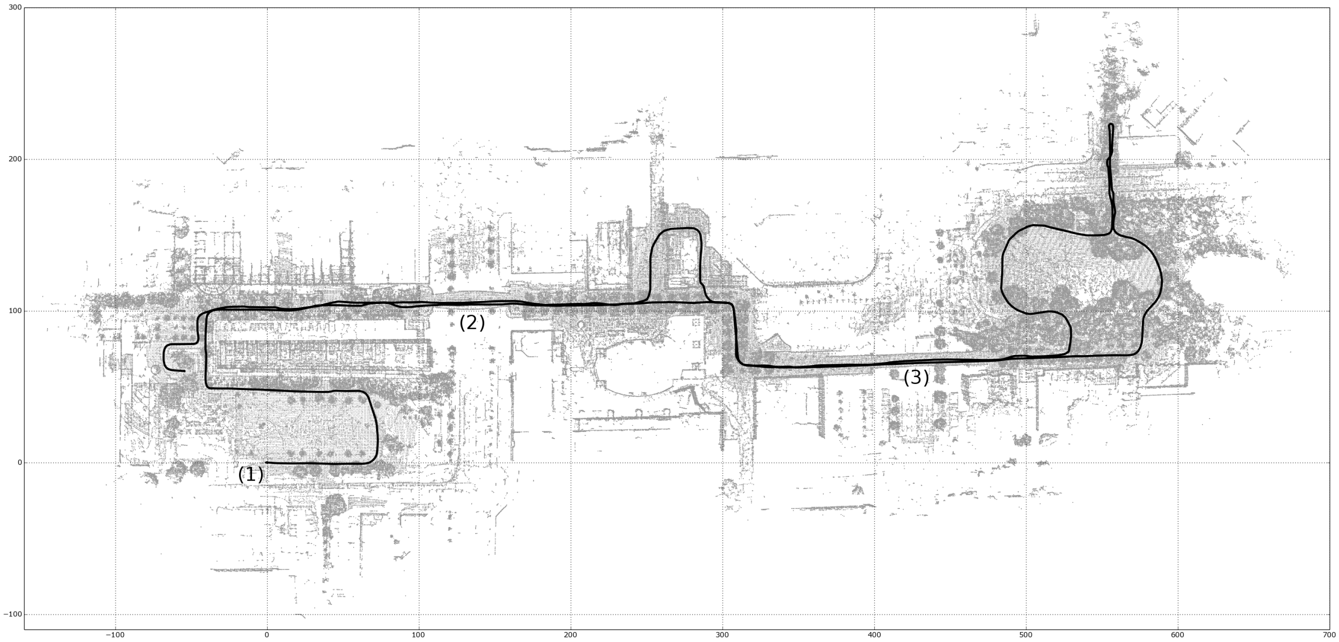


Fig. 7. Trajectory of 2016 Tsukuba Challenge with overlay of top-down point cloud map projection. (1) is starting point; (2) and (3) are the bridge area.

Second experiment involved previous (2015) Tsukuba Challenge dataset, on which the results has been reported in our paper. For this experiment, one mapping run and two localization runs were performed within same timeframe. Similar to 2016 experiment, ground truths were established from LIDAR-based measurements.

Third experiment is by using map from 2015 dataset but applied to 2016 dataset. The objective of this experiment is to find out if our vision-based localization method is applicable for lifelong usage.

5.1. Data Collection

To evaluate our method, we collected two types of datasets, which consists of 2015 and 2016 Tsukuba Challenge track. All datasets consist of image streams from PointGrey Grasshopper3 camera and LIDAR scans from Velodyne HDL-32. The LIDAR scans were used for establishing ground truths in both mapping and localization.

The robot platforms for each year were different but system architecture related to localization is almost all same. The platform for 2015 Tsukuba Challenge has been described in previous paper. For 2016, we used the robot shown in **Fig. 8** and G-Tune as laptop (CPU: Intel Core i7-6700HQ, RAM 64 GB, GPU: GeForce GTX 970M). Although G-Tune installs a GPU, we only used CPU to perform proposed vision-based localization.

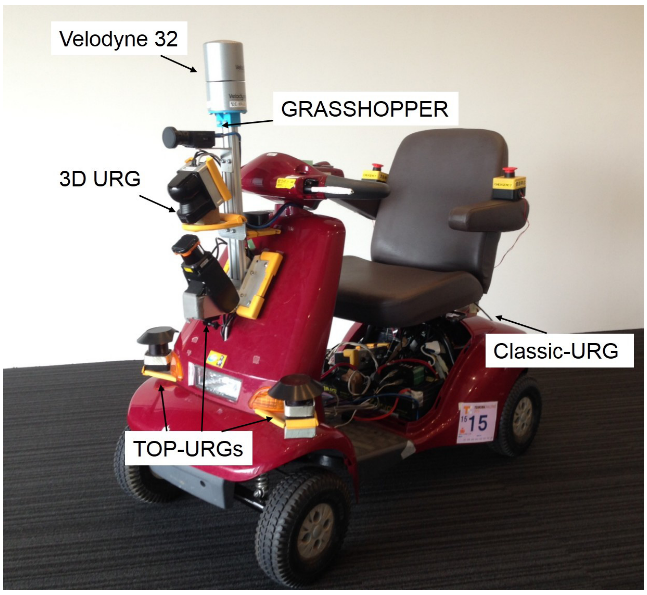


Fig. 8. Experimental platform used in Tsukuba Challenge 2016.

5.2. Evaluation Criteria

In this paper, we use three criteria to evaluate performance of our localization method. First, *coverage* measures percentage of time the robot was able to localize its position in the map. Second, *accuracy* measures metric error of robot pose estimation related to ground truth from NDT localization.

6. Experiment Results and Discussions

The results of our three experiments are summarized in **Table 1** for 2015 experiment, **Table 2** for 2016 experiment, and **Table 3** for long-term localization experiment. In general, our method shows improvements in term of coverage; previously, our method recorded cover-

Table 1. Coverage and accuracy from 2015 experiment.

Runs	Coverage [%]		Current Errors [m]			Prev. Errors [m] [7]		
	Current	Previous	Avg.	Max.	St.Dev	Avg.	Max.	St.Dev
11-03	96.7	68.3	0.74	13.36	0.91	0.38	26.41	1.60
11-07	97.3	68.4	0.68	3.08	0.49	0.08	1.21	0.11

Table 2. Coverage and accuracy from 2016 experiment.

Runs	Coverage [%]	Errors [m]		
		Avg.	Max.	St.Dev.
10-15 13:56	90.8	0.13	3.05	0.10
10-15 15:14	98.5	0.14	1.86	0.12
10-16 13:32	97.2	0.16	3.72	0.15
10-16 14:36	98.4	0.16	2.21	0.17

Table 3. Coverage of localization in 2016 experiment using 2015 map.

Datasets	Coverage (cross-track) [%]
10-16 13:56	19.7 (75.8)
10-16 15:14	16.7 (64.0)
10-16 13:32	25.7 (98.5)
10-16 14:36	16.5 (64.0)

age about 68% in 2015 datasets using single map. With current modification, our single-map vision-based localization shows high percentage (90% at minimum) when using maps created from corresponding date and time (i.e., same year).

Table 3 shows coverage performance of our vision-based localization in 2016 experiment using map created from 2015 as a type of lifelong localization. Overall, map created from last previous year does not perform well due to low overlap between each trajectory (26.1%). However, relative comparison in only overlap areas results in favorable results of lifelong localization. Due to significantly different ground truths between two years, we could not report on accuracy of lifelong localization.

6.1. Coverage

As shown in **Table 1**, the coverage for first experiment (2015 datasets) shows high level of coverage; 96.7% and 97.3% for first and second runs respectively. Compared to our previous results, there have been significant improvements in term of localization coverage; our previous paper recorded coverage of 67%. In current results, coverage improves to more than 97%. This means that the vision-based localization has been quicker to recover from the lost occurrence. The potential areas of lost are shown in **Fig. 9**. We found that these lost events were mostly took place either in the turnings or strong intensities (and smears).

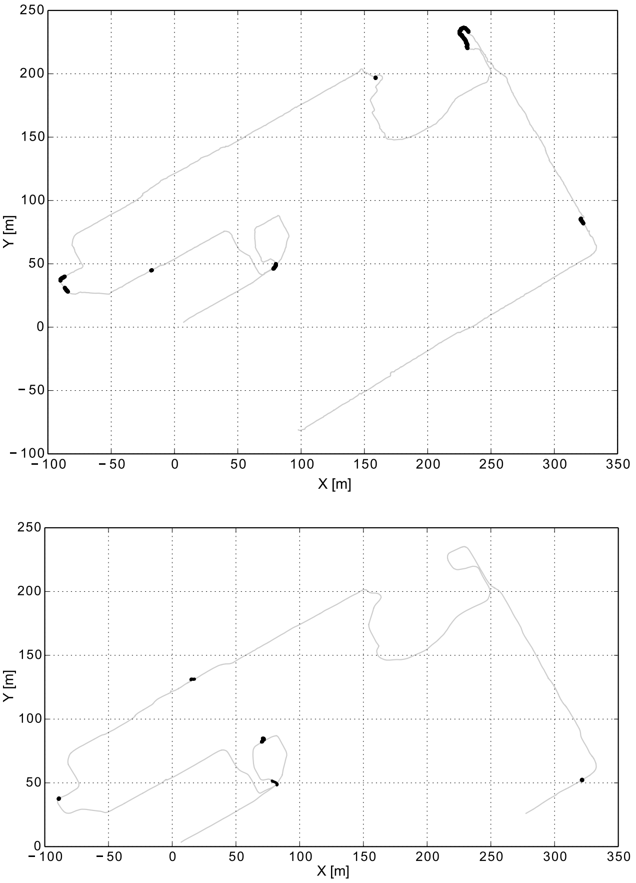


Fig. 9. Lost events position according to ground truth in the 2015 datasets experiment as pointed by bold points. Top: lost occurrences for 2015-11-03. Bottom: 2015-11-07.

For second experiment, our vision-based localization method showed more variations, but still exhibited high level of coverage. Lowest coverage came from first run of 2016 (15th October 13:56) that amounts to 91%. In this dataset, as shown in **Fig. 10**, we encountered long part of lost occurrence that happened after hard bump prior to entering the bridge. Other significant part of unrecoverable vision localization was in the forest area, in which the camera was facing the sun, thus getting frequent lens smears.

In the third experiment, we performed localization in the 2016 datasets using map created from 2015 (**Table 3**). Overall, 2015 map could only covered less than 26% of the 2016 track. This is understandable, since only 26.1% of 2016 trajectory that overlaps with 2015 one.

A particular part of Tsukuba Challenge that is difficult for lifelong localization is the paved pedestrian area cov-

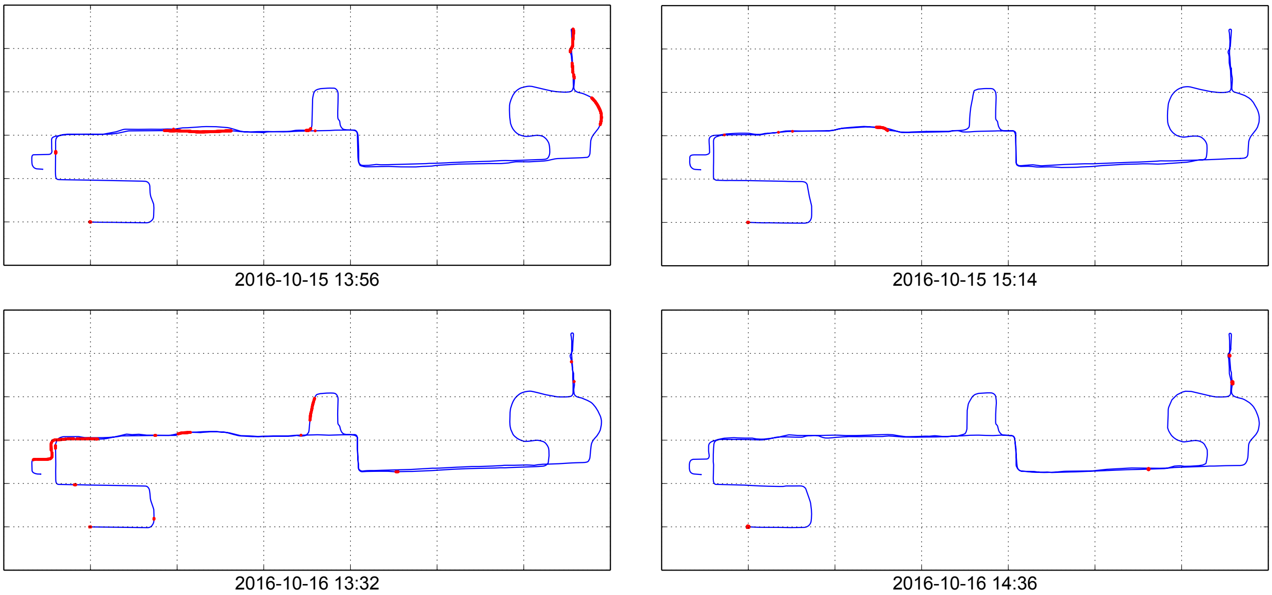


Fig. 10. Positions of lost occurrences in the 2016 datasets as pointed by red markers.

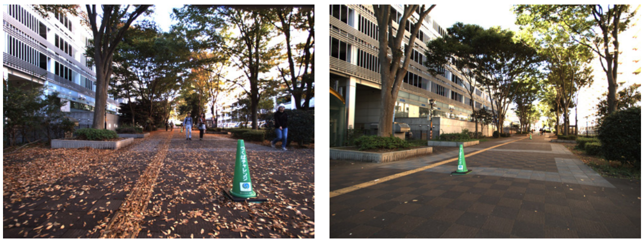


Fig. 11. Same place, unrecognized: left is 2015 situation, while right is from 2016 in the same place.

ered by large trees. In 2015 datasets, large amount of this area were littered by falling leaves; however this cover was almost non-existent in 2016. Therefore, the changes between both years were substantial; making the place recognition subsystem failed. **Fig. 11** shows an example of this situation. Inversely, prominent places where static image features are dominant and highly visible make the place recognition easier. Example of these places are the starting point and the bridge area (pointed by (1) and (2) in **Fig. 7**). Overall, coverages for third experiment are shown in **Fig. 12**.

6.2. Accuracy

Tables 1 and **2** list errors of our vision-based localization method in all experiments. For the 2015 datasets, the new method recorded lower accuracy than previous one, as shown by the average errors for both testing runs. However, in the first test runs of 2015 dataset the large maximum errors was improved to 13.4 m from previous error of 26.4 m. This improvement did not take place in second run, as maximum error had increased to 3.1 m.

Our method registered much better accuracy in the 2016 experiment. The maximum average errors is now below 20 cm, while the maximum errors are significantly reduced below 4 m. Also, overall maximum errors has dropped significantly below in order of below 4 m.

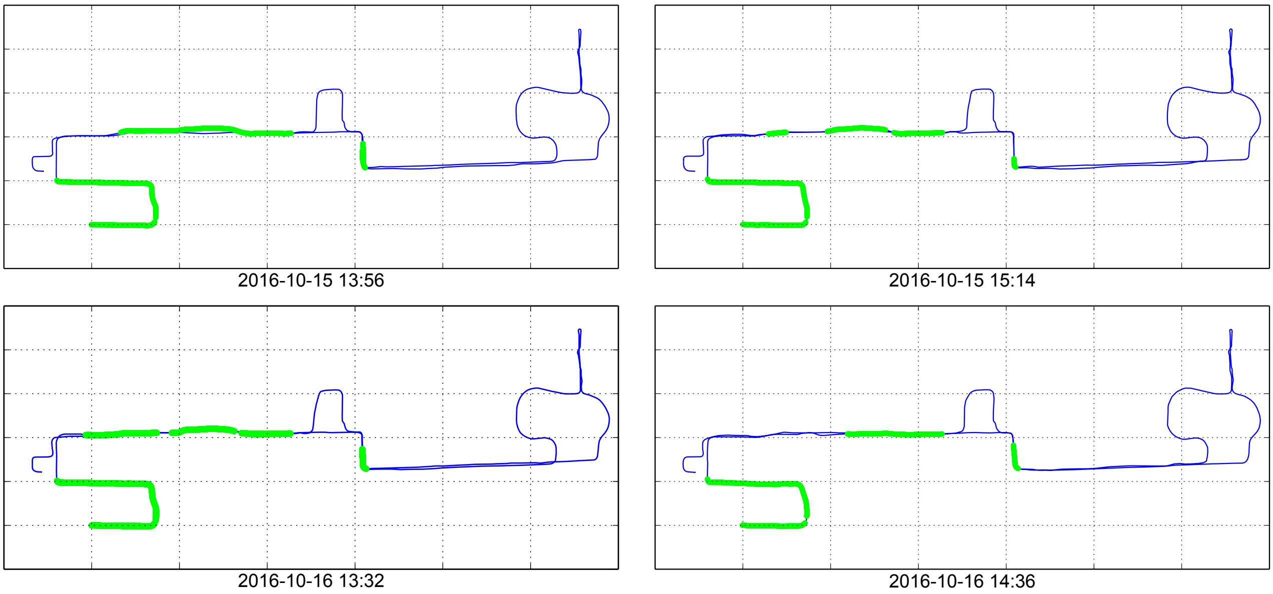
To identify sources of localization errors and how they develop, the size of error are plotted as circles in their respective locations for each experiments. For the sake of brevity, only one datasets are plotted from each experiments as all of the datasets behave similarly in term of error distribution. This error distribution relative to locations are plotted in **Figs. 13** and **14**. From both experiments, most of large errors occurred in three location types:

- Before and/or after recovery from lost.
- Hard turns.
- Open space, where image features fall in far places.

Relationships between accuracy and coverage for all datasets are shown in **Fig. 15**. Here, for 2016 experiment most of the time (above 90%) localization system were able to provide positions within errors below 50 cm, which is adequate for most purpose of navigation. For the rest of time, sensor fusion with odometry will be able to cover the localization requirement [7]. This sensor fusion is also able to mask the large "jumps" that occasionally appears. For 2015 experiment, during 80% of time the localization system could only provide accuracy within 1.2 m, which is not enough for navigation. This was caused by time discrepancies between computers used in logging the image stream and LIDAR.

6.3. Computational Time

Figure 16 plotted fluctuation of per-frame computational time of typical localization test. In average, per-frame time amounts to 83.4 milliseconds, that equals to 12 frame per seconds. This is slightly lower to 15 fps of


2016-10-15 13:56

2016-10-15 15:14

2016-10-16 13:32

2016-10-16 14:36

Fig. 12. Coverage plots of 2016 track using 2015 map; green markers point to navigable areas.

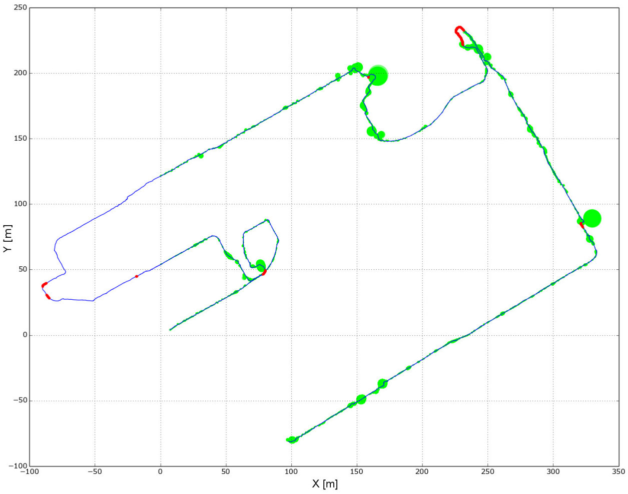


Fig. 13. Error distribution by position for 2015 experiment. Lost occurrences are marked by red points.

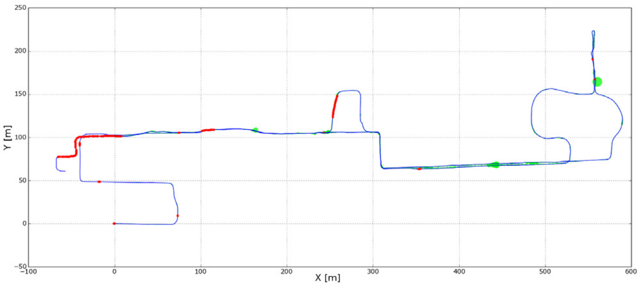


Fig. 14. Error distribution by position for 2016 experiment. Lost occurrences are marked by red points.

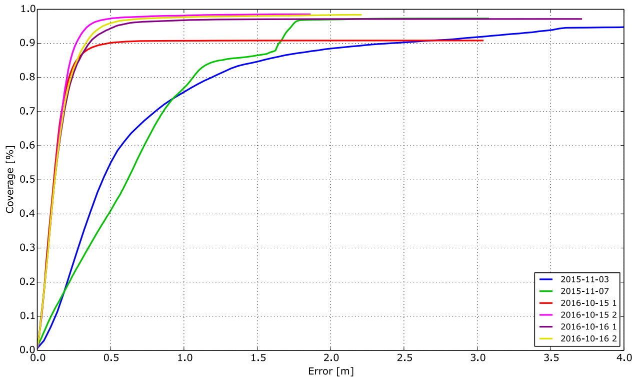


Fig. 15. Cumulative distribution function of errors in each dataset.

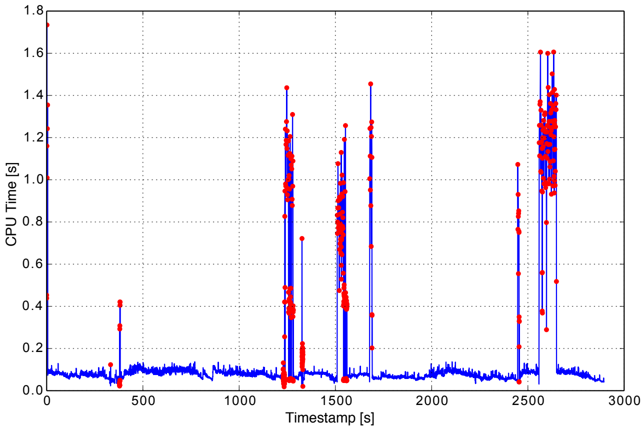


Fig. 16. Fluctuation of CPU time per frame; lost occurrences are marked by red points.

camera image rate that we use, but usable for real-time (as comparison, typical LIDAR-based localization methods make for 10 Hz due to hardware scan rate). However, when localization is lost, the system will perform global relocalization by place recognition that increase computation time significantly (occasionally exceeding 1 second). Compared to the original ORB-SLAM, these surges of CPU usage are not good signs for real-time usage. The increase of time amount is proportional to number of keyframe candidates for place recognition.

6.4. Discussions

From these three experiments and by looking at three parameters (coverage, accuracy and computational time), we put together the following findings:

1. *Modified keyframe search for place recognition with custom vocabulary work correctly*. Compared to our previous results, the localization system has successfully addressed lack of coverage. In the 2015 datasets experiment and 2016 datasets experiment as shown in **Figs. 9** and **10**, lost occurrences have dropped and the system is now able to recover quickly.

2. *Image features from both far and near places are required*. Features from prominent landmarks (i.e., buildings) could help for place recognition. However, existence of this type of features alone without features from near places may cause visual odometry subsystem to deduce very small motion. In contrast, near-place features (e.g., from trees and paving blocks) are not quite useful for landmarks as shown by third experiment because they are prone to changes. In this regard, gamma control to regain brightness in dark areas has contribution for accuracy of localization as it could help to recover features from near places.

3. *Lifelong localization is possible*. In the third experiment, localization performed successfully in the places that had not change considerably. Also, as shown from second findings above, prominent landmarks in the frame will help for global localization.

4. *There is trade-off between robust place recognition and CPU usage*. Compared to the original ORB-SLAM, our keyframe search method basically performs brute-force search against all candidates rather than filters just the most likely ones. As a consequence, this increases CPU time as complexity of scoring function of keyframe matches is linear against number of features.

7. Conclusions and Future Work

In this work, we have described and evaluated our vision-based localization in the Tsukuba Challenge environments. From the point of view of coverage, the localization system is capable to provide high availability, higher than 90% for all the log data tested in the Tsukuba track. It also could cope with data of different years providing coverage of 75%, as long as environmental changes were low. The coverage was improved using a combination of non-discriminatory keyframe selection and custom vocabulary that is unique for each scene.

The capability to support fully vision-based navigation is still limited due to concern of large errors that may occur in some areas. We have shown previously that it is possible to combine this method with other metric localization systems such as odometry with particle filter which can cope with these issues. It is left for future work the exploration for methods that strive to increase accuracy for any general situations.

We would like to explore the combination of LIDAR and camera in the mapping process. Especially, we would like to explore and eliminate pose estimation by camera and replace it using accurate pose from NDT. By doing this, triangulation process will result in metric landmarks and map points. This will lead to the elimination of scale correction as pose estimation has already been in metric. Another effect is that particle filter may be evaluated directly in metric coordinate, as particle position is now able to use map point projection for scoring.

Another area worth investigating is *Long-Term Mapping*; which would provide with the capability to build and grow a map of the area from different times. In this regard, we have proved possibility of lifelong localization but also noted failures in the unmapped or vastly changing areas.

Acknowledgements

This work is supported by MEXT COI stream program, A Diverse and Individualized Social Innovation Hub – The "Mobility Society" for the Elderly: Lead to an Active and Joyful Lifestyle. The robot for this experiment is provided by Tetsuo Tomizawa et al. Our fork of ORB-SLAM is available as part of Autoware, open source software for urban autonomous driving [36].

References:

[1] M. Milford, "Vision-based place recognition: how low can you go?," The Int. J. of Robotics Research, Vol.32, No.7, pp. 766-789, 2013.

[2] M. Buerki, I. Gilitschenski, E. Stumm, R. Siegwart, and J. Nieto, "Appearance-Based Landmark Selection for Efficient Long-Term Visual Localization," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), Daejeon, Korea, October 2015.

[3] C. Linegar, W. Churchill, and P. Newman, "Work Smart, Not Hard: Recalling Relevant Experiences for Vast-Scale but Time-Constrained Localisation," Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), Seattle, WA, USA, May 2015.

[4] C. Valgren and A. J. Lilienthal, "SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments," Robotics and Autonomous Systems, Vol.58, No.2, pp. 149-156, 2010.

[5] K. Irie, T. Yoshida, and M. Tomono, "Outdoor Localization Using Stereo Vision Under Various Illumination Conditions," Advanced Robotics, Vol.26, No.3-4, pp. 327-348, 2012.

[6] N. Akai, K. Yamauchi, and K. Inoue, "Development of Mobile Robot "SARA" that Completed Mission in Real World Robot Challenge 2014," Special Issue on Real World Robot Challenge in Tsukuba: Autonomous Technology for Useful Mobile Robot, J. of Robotics and Mechatronics, Vol.27, No.4, pp. 327-336, Aug. 2015.

[7] A. Sujiwo, T. Ando, E. Takeuchi, Y. Ninomiya, and M. Edahiro, "Monocular Vision-Based Localization Using ORB-SLAM with LIDAR-Aided Mapping in Real-World Robot Challenge," Special Issue on Real World Robot Challenge in Tsukuba: Autonomous Technology for Coexistence with Human Beings, J. of Robotics and Mechatronics, Vol.28, No.4, pp. 479-490, 2016.

[8] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a Versatile and Accurate Monocular SLAM System," IEEE Trans. on Robotics, 2015.

[9] E. Takeuchi and T. Tsubouchi, "A 3-D Scan Matching using Improved 3-D Normal Distributions Transform for Mobile Robotic Mapping," 2006 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 3068-3073, 2006.

[10] R. Kümmerle, M. Ruhnke, B. Steder, C. Stachniss, and W. Burgard, "Autonomous Robot Navigation in Highly Populated Pedestrian Zones," J. of Field Robotics, 2014.

[11] Y. Morales, A. Carballo, E. Takeuchi, A. Aburadani, and T. Tsubouchi, "Autonomous robot navigation in outdoor cluttered pedestrian walkways," J. of Field Robotics, Vol.26, No.8, pp. 609-635, 2009.

[12] K. Konolige and J. Bowman, "Towards lifelong visual maps," 2009 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 1156-1163, Oct. 2009.

[13] P. Mühlfellner, M. Bürki, M. Bosse, W. Derendarz, R. Philippsen, and P. Furgale, "Summary Maps for Lifelong Visual Localization," J. of Field Robotics, Vol.33, No.5, pp. 561-590, 2016.

[14] B. Paden, M. Cap, S. Z. Yong, D. Yershov, and E. Frazzoli, "A Survey of Motion Planning and Control Techniques for Self-driving Urban Vehicles," IEEE Trans. on Intelligent Vehicles, 2016.

[15] H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Scale Drift-Aware Large Scale Monocular SLAM," Robotics: Science and Systems, Vol.2, p. 5, 2010.

[16] R. Hartley and A. Zisserman, "Multiple view geometry in computer vision," Cambridge university press, 2003.

[17] C. Siagian and L. Itti, "Biologically Inspired Mobile Robot Vision Localization," IEEE Trans. on Robotics, Vol.25, No.4, pp. 861-873, 2009.

[18] C.-K. Chang, C. Siagian, and L. Itti, "Mobile robot vision navigation & localization using gist and saliency," 2010 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 4147-4154, 2010.

[19] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard, "Monocular Camera Localization in 3D LiDAR Maps," 2016 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), October 9, 2016.

[20] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual Place Recognition: A Survey," IEEE Trans. on Robotics, Vol.32, No.1, pp. 1-19, 2016.

[21] C. Kanan and G. W. Cottrell, "Color-to-Grayscale: Does the Method Matter in Image Recognition?," PLOS ONE, Vol.7, No.1, e29740, 2012.

[22] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman, "Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles," Proc. of the Visual Place Recognition in Changing Environments Workshop, IEEE Int. Conf. on Robotics and Automation (ICRA), Hong Kong, China, Vol.2, p. 3, 2014.

[23] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," Artificial Intelligence Review, Vol.43, No.1, pp. 55-81, 2015.

[24] G. Ros, A. Sappa, D. Ponsa, and A. M. Lopez, "Visual slam for driverless cars: A brief survey," Intelligent Vehicles Symposium (IV) Workshops, 2012.

[25] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," 2011 IEEE Int. Conf. on Computer Vision (ICCV), pp. 2564-2571, 2011.

[26] J. Heinly, E. Dunn, and J.-M. Frahm, "Comparative evaluation of binary features," Computer Vision – ECCV 2012, pp. 759-773, Springer, 2012.

[27] M. Lourakis and X. Zabulis, "Accurate scale factor estimation in 3D reconstruction," Computer Analysis of Images and Patterns, pp. 498-506, Springer, 2013.

[28] R. Szeliski and S. B. Kang, "Shape ambiguities in structure from motion," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.19, No.5, pp. 506-512, 1997.

[29] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós, "A comparison of loop closing techniques in monocular SLAM," Robotics and Autonomous Systems, Vol.57, No.12, pp. 1188-1197, 2009.

[30] D. Galvez-López and J. Tardos, "Bags of Binary Words for Fast Place Recognition in Image Sequences," IEEE Trans. on Robotics, Vol.28, No.5, pp. 1188-1197, 2012.

[31] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.31, No.4, pp. 591-606, 2009.

[32] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree," 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'06), Vol.2, pp. 2161-2168, 2006.

[33] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," Proc. Ninth IEEE Int. Conf. on Computer Vision, pp. 1470-1477, Vol.2, Oct. 2003.

[34] A. Adams and R. Baker, "The negative," New York Graphic Society, 1981.

[35] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," 6th IEEE and ACM Int. Symposium on Mixed and Augmented Reality 2007 (ISMAR 2007), pp. 225-234, 2007.

[36] S. Kato, E. Takeuchi, Y. Ishiguro, Y. Ninomiya, K. Takeda, and T. Hamada, "An open approach to autonomous vehicles," IEEE Micro, Vol.35, No.6, pp. 60-68, 2015.

Name:
Adi Sujiwo

Affiliation:
Department of Information Engineering, Graduate School of Informatics, Nagoya University

Address:
609 National Innovation Complex (NIC), Furo-cho, Chikusa-ku, Nagoya
464-8601, Japan

Brief Biographical History:

2009-2011 Magister of Computer Science, University of Indonesia
2011-2014 System Engineer, Bogor Agricultural University, Indonesia
2014- Ph.D. Student, Nagoya University

Name:
Eijiro Takeuchi

Affiliation:
Department of Intelligent Systems, Graduate School of Informatics, Nagoya University

Address:
609 National Innovation Complex (NIC), Furo-cho, Chikusa-ku, Nagoya
464-8601, Japan

Brief Biographical History:

2008-2015 Assistant Professor, Tohoku University
2015-2016 Designated Associate Professor, Nagoya University
2016- Associate Professor, Nagoya University

Main Works:

- "A 3-D Scan Matching using Improved 3-D Normal Distributions Transform for Mobile Robotic Mapping," 2006 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2006), Beijing, China, pp. 3068-3073, 2006.

Membership in Academic Societies:

- The Japan Society of Mechanical Engineering (JSME)
- The Robotics Society of Japan (RSJ)
- The Society of Instrument and Control Engineers (SICE)
- The Institute of Electrical and Electronics Engineers (IEEE)

Name:
Luis Yoichi Morales

Affiliation:
Driving Scene Understanding Research Division, Institute of Innovation for Future Society, Nagoya University

Address:
609 National Innovation Complex (NIC), Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Brief Biographical History:

2006 Received M.Eng., Intelligent Robot Laboratory, Tsukuba University
2009 Received Ph.D., Intelligent Robot Laboratory, Tsukuba University
2010-2016 Researcher, ATR Intelligent Robotics and Communication Laboratories
2016- Designated Associate Professor, Nagoya University

Membership in Academic Societies:

- The Institute of Electrical and Electronics Engineers (IEEE) Robotics and Automation Society
- The Robotics Society of Japan (RSJ)

Name:
Naoki Akai

Affiliation:
Driving Scene Understanding Research Division, Institute of Innovation for Future Society, Nagoya University

Address:
609 National Innovation Complex (NIC), Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Brief Biographical History:

2013-2016 Doctor Student, Utsunomiya University
2016- Designated Assistant Professor, Nagoya University

Main Works:

- "Development of mobile robot "SARA" that completed mission in real world robot challenge 2014," J. of Robotics and Mechatronics, Vol.27, No.4, pp. 327-336, 2015.
- "Gaussian processes for magnetic map-based localization in large-scale indoor environments," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 4459-4464, 2015.
- "Robust localization using 3D NDT scan matching with experimentally determined uncertainty and road marker matching," IEEE Intelligent Vehicles Symposium, pp. 1364-1371, 2017.

Membership in Academic Societies:

- The Institute of Electrical and Electronics Engineers (IEEE)
- The Japan Society of Mechanical Engineering (JSME)
- The Robotics Society of Japan (RSJ)
- Society of Instrumentation and Control Engineering (SICE)

Name:
Hatem Darweesh

Affiliation:
Department of Intelligent Systems, Graduate School of Informatics, Nagoya University

Address:
085 IB North, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Brief Biographical History:

1998-2002 B.Sc., Faculty of Computers and Information Sciences, Ain Shams University
2002-2013 Teaching Assistant, Moder Academy in Maadi
2004-2009 M.Sc., Faculty of Computers and Information Sciences, Ain Shams University
2004-2010 Co-Founder, NRG Solutions, Egypt
2013-2016 Robotics Engineer, ZMP Inc., Japan
2016- Ph.D. Student, Graduate School of Information Science, Nagoya University

Name:
Yoshiki Ninomiya

Affiliation:
Intelligent Vehicle Research Division, Institute of Innovation for Future Society, Nagoya University

Address:
609 National Innovation Complex (NIC), Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Brief Biographical History:

1981 Received B.S. from Nagoya University
1983 Received M.S from Nagoya University
1983-2003 Researcher, Toyota Central Lab.
2008 Received Ph.D. from Nagoya University
2014- Designated Professor, Nagoya University

Name:
Masato Edahiro

Affiliation:
Department of Information Engineering, Graduate School of Informatics, Nagoya University

Address:
4F IB South, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Brief Biographical History:

1985 NEC Research Center
1999 Ph.D. in Computer Science, University of Princeton
2011- Professor, Graduate School of Informatics, Nagoya University