PAPER

Position-Invariant Robust Features for Long-Term Recognition of Dynamic Outdoor Scenes

Aram KAWEWONG[†a)], Sirinart TANGRUAMSUB[†], *Nonmembers*, *and* Osamu HASEGAWA[†,††], *Member*

SUMMARY A novel Position-Invariant Robust Feature, designated as PIRF, is presented to address the problem of highly dynamic scene recognition. The PIRF is obtained by identifying existing local features (*i.e*. SIFT) that have a wide baseline visibility within a place (one place contains more than one sequential images). These wide-baseline visible features are then represented as a single PIRF, which is computed as an average of all descriptors associated with the PIRF. Particularly, PIRFs are robust against highly dynamical changes in scene: a single PIRF can be matched correctly against many features from many dynamical images. This paper also describes an approach to using these features for scene recognition. Recognition proceeds by matching an individual PIRF to a set of features from test images, with subsequent majority voting to identify a place with the highest matched PIRF. The PIRF system is trained and tested on 2000+ outdoor omnidirectional images and on COLD datasets. Despite its simplicity, PIRF offers a markedly better rate of recognition for dynamic outdoor scenes (*ca*. 90%) than the use of other features. Additionally, a robot navigation system based on PIRF (PIRF-Nav) can outperform other incremental topological mapping methods in terms of time (70% less) and memory. The number of PIRFs can be reduced further to reduce the time while retaining high accuracy, which makes it suitable for long-term recognition and localization.

key words: *scene localization, scale invariant feature transformation (SIFT), scene recognition, topological mapping*

1. Introduction

Localization is an indispensable capability for both humans and machines. Knowing "Where we are" has always been an important topic in robotics and computer vision communities. Especially for mobile robots, knowing its position is a fundamental requirement for navigation systems. The topic has been studied for more than two decades using several methods: metrical, topological, and hybrid (see [1], [2] for reviews). Although sonar and laser scanners have traditionally been the sensory modalities of choice [3], current advances in visual tools have made visual approaches more attractive, providing richer information at a lower price.

From the perspective of computer vision, an efficient robot vision system might need to overcome three difficulties: dynamical changes, viewpoint changes, and scene categorization. In a highly dynamic environment (i), places might look very different over time because of illumination changes (daytime, nighttime) and because of moved objects: parking lots are empty on holidays. These changes are dynamic because their appearances are stable only for only some period of time. Regarding the second sub-problem (ii), different viewpoints often make a scene look different. This problem also includes changes in weather and lighting conditions. An object's appearance might be very different when observed from different camera positions, even if viewed at exactly same time. Scene categorization (ii) describes how the robot understands the scene so that it can categorize new unseen places along with those it has seen previously. Inspired by biology, this ability further reduces the gap separating robots and humans. Recent scene recognition approaches might be divided into three main types: *Object-Based*, *Region-Based*, and *Context-Based*.

To date, most approaches to scene recognition have been *object-based* [4]–[6]. Using such approaches, a scene location is recognized by identifying a set of landmarks known to be included in a scene. These approaches are prone to carrying over and amplifying low-level errors along the stream of processing. For instance, upstream identification of small objects (pixel-wise) is hindered by the downstream noise inherent to camera sensors and by varied lighting conditions. This is problematic in spacious environments where landmarks are more dispersed and more distant from the agent. This approach must be environment-specific to ensure the simplicity of selecting a small set of anchor objects as landmarks in an open problem.

For *region-based* scene recognition, the segmented image regions and their configurational relations are used to form a signature of a location. The major problem hindering this approach is reliable region-based segmentation, in which individual regions must be characterized robustly and associated. Naïve template matching involving a rigid relation is often insufficiently flexible in the face of undersegmentation or over-segmentation, which is often true with unconstrained environments, such as outdoors. Some techniques such as normalized-cut [17], [35] are useful to improve segmentation quality. Nevertheless, the computation time for image segmentation might be problematic for real time applications.

Context-based approaches, unlike both previously described approaches, bypass traditional processing steps. Context-based approaches examine the input image as a whole and extract a low-dimensional signature that compactly summarizes the image's statistics and semantics. The challenge of discovering a compact and holistic representa-

Manuscript received December 17, 2009.
Manuscript revised April 8, 2010.

[†]The authors are with the Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama-shi, 226–8503 Japan.

[††]The author is with Imaging Science and Engineering Laboratory, Tokyo Institute of Technology, Yokohama-shi, 226–8503 Japan.

a) E-mail: kawewong.a.aa@m.titech.ac.jp

DOI: 10.1587/transinf.E93.D.2587

tion for unconstrained images has therefore prompted considerable research effort recently. Renniger and Malik [18] use a set of texture descriptors and a histogram to create an overall profile of an image. Ulrich and Nourbakhsh [19] build color histograms and perform matching using a voting classifier. Oliva and Torralba [20] encode some spatial information by performing 2D Fourier Transform analyses in individual image subregions on a regularly spaced grid. The resulting spatially arranged set of signatures, one per grid region, is further reduced using principal component analysis (PCA) to yield a unique low-dimensional image classification. In more recent implementations, Torralba et al. [21] used steerable wavelet pyramids instead of the Fourier Transform to solve the robotic localization.

Among the three approaches described above, PIRF is most related to an *object-based* approach, in the sense that natural landmarks would be used as a signature of place. Our selection is motivated by the observation that outdoor scenes generally include distant objects such as distant buildings or walls. These distant objects seem to appear constantly in scenes irrespective of the camera position. Even in a highly dynamic scene where major components of scenes are changed, these small distant objects still appear. In this case, global representation of whole scenes might be problematic; such scenes include many unstable nearby objects whose subsequent recognition fails. Therefore, we address this highly dynamic scene recognition problem as:

1. how to detect objects *autonomously* which are visible to almost every position in such place;
2. how to detect objects autonomously which are unique to a single place; and
3. how to describe such objects precisely despite their distance.

These distant objects can be a good signature of each place; a group of these objects can be used efficiently to identify places. As described herein, we propose a Position-Invariant Robust Feature (PIRF) as an image local feature that solves these three problems.

The PIRF is developed upon existing local descriptors such as Scale-Invariant Feature Transformation (SIFT) [7] and Speeded Up Robust Feature (SURF) [13]. Local features extracted from many individual images are filtered to derive the descriptors which appear repeatedly in almost every scene (taken at the same place). These descriptors are averaged to generate a single representative descriptor called PIRF. Filtering is done autonomously using simple feature matching as in an earlier study [7]. Considering Fig. 1, given several SIFTs extracted from many interesting points of an image (left), and assuming that some SIFTs appear repeatedly in a few more sequential images, PIRFs are then generated by interpolating those corresponding SIFTs. Figure 1 (right) shows PIRFs extracted from standard SIFTs. In this case, almost all PIRFs are only of distant objects, whose appearances are very stable. This solves the *first* problem. One place can be represented using several PIRFs because one place might include many distant objects; each requires many PIRFs for representation. For later reference in this paper, we call these representative PIRFs (for one individual place) a PIRF-dictionary. For example, three places require three PIRF-dictionaries for representation. Because of the descriptive power of the existing descriptor, detected objects can be described precisely using a set of PIRFs (representatives of slow-moving SIFTs or SURFs). This solves the *third* problem described earlier. However, it is also clear that collecting many PIRFs will eventually pose the problem of many duplicated PIRFs, which confuse recognition in the long run. Therefore, we additionally propose a technique for eliminating these redundant PIRFs. The technique is incremental; at any time, it can rapidly search for redundant PIRFs and delete them from memory. This solves the *second* problem above.

We also describe a simple approach to use PIRFs for scene recognition. The recognition system is portrayed in Fig. 2. Assume that an environment contains five separate places. Each place has its own PIRF-dictionary $\mathcal{D}^i, i \leq 5$ for representation. First, a set of SIFTs is extracted from a testing image. Feature matching is performed to match each single SIFT to a set of PIRFs in each dictionary. A place, whose dictionary contains the highest number of PIRFs that can be matched to the extracted SIFTs, is justified as the winner. In Fig. 2, for instance, both the first and the second image belong to place 1 because the number of matches between SIFTs and PIRFs in $\mathcal{D}^1$ is the highest.

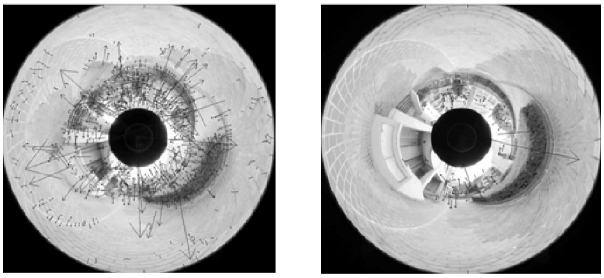


Fig. 1 Sample omnidirectional outdoor image extracted with original SIFT (left) and the proposed PIRF (right). The distant objects' appearances are invariant to position changes.

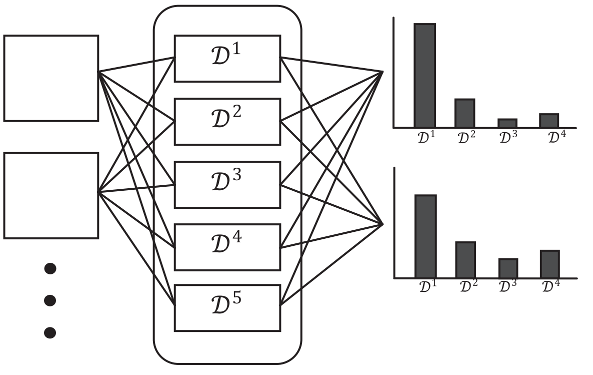

$\mathcal{D}^1$
$\mathcal{D}^2$
$\mathcal{D}^3$
$\mathcal{D}^4$
$\mathcal{D}^5$
$\mathcal{D}^1$ $\mathcal{D}^2$ $\mathcal{D}^3$ $\mathcal{D}^4$
$\mathcal{D}^1$ $\mathcal{D}^2$ $\mathcal{D}^3$ $\mathcal{D}^4$

Fig. 2 Illustration of PIRF voting for scene recognition. Each dictionary votes for the matched PIRF included in an input image. The voting result justifies the location of the scene.

To demonstrate the advantages of PIRF, we test it on 1000+ outdoor omnidirectional images collected from two campuses. Training and testing data are collected, respectively, on holidays and weekdays to address the difficulties of highly dynamic changes over time. The results show that PIRF obtains a markedly higher rate of recognition than other features.

We also describe a simple robot navigation method that uses PIRF (designated as PIRF-Nav) as a basic feature, to confirm PIRF's importance in relation to robotics. The PIRF-Nav is tested on one campus. The image dataset is identical to that used previously for scene recognition. Results show that PIRF-Nav outperforms Incremental Spectral Clustering [15], [16] in both time and accuracy.

2. Related Works

Various approaches used in the past have addressed Scene and Place recognition. Many effective features and various modes of use have been proposed. Histograms of image properties, *e.g.* color [19] or image derivatives have been used widely in place recognition. However, after SIFT [7] was popularized among the vision community, it came to dominate feature choice in place recognition systems [15], [22], [23]. The SIFT features are invariant to scale and are robust to rotation changes. The 128 dimensional SIFT descriptors have high discriminative power, but are simultaneously robust to local variations [39]. For place recognition, SIFT outperforms edge points [9], pixel intensities [38], and steerable pyramids [37].

As the most appealing descriptor for practical uses, SIFT has been used widely in appearance-based navigation [10], [11] because matching digital image contents among different views of scene requires a good distinctive invariant feature. Although SIFT satisfies such a requirement, its computation time and memory cost are high. As a remedy, Ledwich and Williams [12] reduced SIFT features by taking advantage of the structure of the indoor environment where average view-depths of most images are short. This makes vertical planes such as walls dominate an image's composition. The rotational invariance of SIFT has also been removed by assuming that the viewpoint for indoor images will be stable to rotation around the view axis, resulting in a non-rotated orientation of the keypoint descriptors located on vertical surfaces. The method is specifically useful for indoor environments.

Meanwhile, Oliva and Torralba [20] suggested that recognition of scenes can be achieved using "global configurations", without detailed object information. Consequently, statistical analysis of SIFT distribution became popular. Torralba et al. [21] use global image features to generate Gaussian Mixture Models for place recognition, using fixed variance. The method gives limited tolerance for appearance variation and is not invariant to translation or scale changes. Lazebnik et al. [9] use the *k*-means algorithm to cluster SIFT features, and cluster centers were used as the codebook to solve 15-class scene recognition. Cummins and Newman [29] integrate bag-of-visual-words (BoW) into the recursive probabilistic Bayesian framework and achieve performance beyond the localization. That method can determine that a new image has come from a previously unseen place. Later, Angeli et al. [23] proposed incremental BoW. Starting from an empty dictionary, the system can gradually collect new words while localizing places. Recently, Wu and Rehg [36] proposed the spatial Principle Component Analysis on Census Transform (sPACT) as a feature for scene recognition and categorization. Its performance has proven to be better than that of the BoW method [9]. The authors reported the highest accuracy over the KTH-IDOL dataset [31].

In robotics, scene recognition and localization are also important topics. Several previously proposed methods should also be acknowledged here. Robot localization can be done based on vision alone [19], [24], [25], [30], or based on combinations of vision and other sensors (*e.g.*, laser-scanner, odometer) [19], [26]–[29]. The methods can be performed either offline or online for use indoors or outdoors. Clemente et al. [26] generated metric maps based only on a hand-held camera. The key ingredient of the method is the inverse-depth representation [34], which can estimate the depth of local features even in a single observation. The camera motion is an important concern for representing the main map. Royer et al. [25] reported their success in building the 3D reconstruction of map from sequential scenes by calculating the robot motion. Distinct from [25] and [34], the method of Luo [32] emphasizes single-image recognition: neither robot motions nor its position is incorporated in the method. The incremental support vector machine (SVM) is used to train the localization system for indoor use. The authors apply two techniques to extend SVM to an incremental version: *Fixed Partition* and *Error-Driven*. Nonetheless, the incremental step used in their work contains some images partitioned by the user. The method must go offline until the step has been completed. In addition, the adopted error-driven technique requires a human to stand by the robot's side and tell it whether the recognition is corrected or not. Valgren and Lilienthal [15], [16] use incremental spectral clustering (ISC) to cluster images and thereby create a topological map. Their method was reported as a fully incremental mapping method in highly dynamic outdoor environments (across seasons). The number of nodes in the map (data segmentation) is determined autonomously. The incremental BoW [23] is also a fully incremental mapping method. An empty dictionary can be updated incrementally. This method is considered as state-of-the-art vision-based mapping. However, the method addresses only the loop-closure detection problem; all obtained images up to the current time are included in the process of probabilistic decision-making. They did not report the recognition rate of a single image. In other words, their method is especially applicable for robotics. Furthermore, they did not address the problem of highly dynamic changes in scenes; all images in the study seem to be collected on one day.

Here we confirm again that PIRF is new. From a computer vision perspective, PIRF is more discriminative than global features, but less noisy than using standard local features. Unlike other object-based approaches, PIRF can autonomously detect the landmarks which are robust against highly dynamic changes of scenes. From a robotic perspective, navigation based on PIRF (PIRF-Nav) outperforms ISC in terms of time and accuracy. PIRF-Nav is incremental like [23]; it can recognize single images and close the loops, while its performance is beyond localization like the method of [29]; it can determine that a new observation come from a previously unseen places. Its robustness against dynamic change also leaves some room for additional improvement and applications (*i.e.*, using PIRF as the local feature for generating BoW).

3. PIRF Definition

A Position-Invariant Robust Feature (PIRF) is a single local feature that is robust to any position along the path within the same place. The idea comes from observing that outdoor scenes generally include faraway objects. These objects are useful to identify the place because their appearance is stable, irrespective of position changes. Precisely, PIRF is a single local descriptor computed as the average of existing local descriptors, such as SIFT [7] or SURF [13], which has wide baseline visibility. Actually, a PIRF must be extracted from sequential images because it must retrieve all associated features from these images and compress them into a single PIRF. Many single PIRFs are collected to form an individual PIRF-dictionary of a place (one place contains many sequential images). An individual dictionary is a signature of an individual place. In other words, a video sequence is segmented into N places (partitions). Each place i contains n_i sequential images. It seems very difficult to find a fine number of features that appear in all images. Therefore, the place is divided into many sub-places before extracting PIRF. Figure 3 portrays extraction of PIRFs from a single place. By repeating this extraction process for every place, N dictionaries can be obtained, where each contains PIRFs used for representing an individual place. The algorithm has three main stages: *Sequential Image Matching, PIRF Extraction, and Place Recognition.*

3.1 Sequential Image Matching

Given N as the current number of all visited places in an environment, n_i is the number of members of the sequential image set $\mathbb{I}_i = \{I_1, \ldots, I_{n_i}\}$ of the i^{th} place, where $i \leq N$, I_q is the q^{th} image in the set, $q < n_i$. Each image is described by the 128-D SIFT feature extracted by the standard SIFT algorithm [7]. SIFT matching is performed sequentially for every pair of images; namely $(I_1 - I_2), \ldots, (I_{n_i-1} - I_{n_i})$. We use the same matching criteria as that used in an earlier study [7]. The threshold value is set to 0.6. After every pair of images has been matched, the matching result is retained as the matching index vector of the i^{th} place, $\overrightarrow{m}^i_q = \left(m^i_{1,q}, \ldots, m^i_{k_q,q}\right)$, where $1 \leq q < n_i$, k_q is the index number of local features of image I_{q+1} of the i^{th} place. That is, $m^i_{k_q,q}$ is the integer indicating the index of the local feature in image I_q that match to the $(k_q)^{th}$ feature in image I_{q+1}. For example, $\overrightarrow{m}^1_1 = (10, 0)$ is interpreted as the first matching between I_1 and I_2 of the i^{th} place results in, out of two features, only one matched feature. The first feature of I_1 matches the tenth feature of I_2, whereas the second features of I_1 are not found in the image I_2. As described herein, we select the SIFT of [7] as our descriptor.

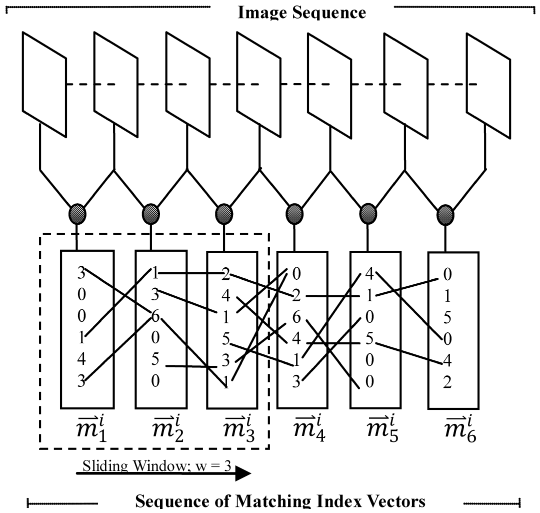


Fig. 3 Sample PIRF extraction of the i^{th} place. Given the number of sequential images $n_i = 7$ and the size of sliding window $w = 3$. Number of extracted SIFT from each image is 6. Every image pair is compared using feature matching, resulting in six matching index vectors. A vector element is the index of the corresponding feature in the next image. For example, for the first sub-place ($\overrightarrow{m}^i_1, \overrightarrow{m}^i_2, \overrightarrow{m}^i_3$) of I_1, I_2, I_3, I_4, there are only three features appearing in all images: $(1, 3, 6, 1)$, $(4, 1, 1, 2)$, $(6, 3, 6, 1)$. $(1, 3, 6, 1)$ is interpreted, respectively, as the 1st, 3rd, 6th, and 1st feature of image I_1, I_2, I_3, I_4. These four features are interpolated to obtain a single representative PIRF. Therefore, there would be 3, 4, 4, 3 PIRFs for the 1st, 2nd, 3rd, and 4th sub-place respectively, 14 PIRFs in all for the whole i^{th} place.

3.2 PIRF Extraction

Considering the $(n_i)^{th}$ image (the last image of the i^{th} place), after $n_i - 1$ matching index vectors $\overrightarrow{m}$ are derived, then the PIRF is extracted. However, an object with a stable appearance irrespective of the changed position is difficult to find because the path might be long or curved. Therefore, we instead extract those features which are positionally invariant in relation to the sub-place. Considering the sequence of vector $\overrightarrow{m}^i_q$ as the sequential input data, sliding windows feature extraction is performed to collect PIRFs from many sub-places instead of the whole place. For example, if $w = 3$, then the first sub-place contains $\overrightarrow{m}^i_1$, $\overrightarrow{m}^i_2$, $\overrightarrow{m}^i_3$ corresponding to I_1, I_2, I_3, I_4, and the second sub-place contains $\overrightarrow{m}^i_2, \overrightarrow{m}^i_3, \overrightarrow{m}^i_4$ corresponding to I_2, I_3, I_4, I_5. The window size is w; the window is shifted by one, which means that, given D^i_j as the PIRF-dictionary containing a set of de-

Algorithm 1: PIRF Extraction of the i^{th} place

```
Require: $\overrightarrow{m}$ is the matching index vector, as shown in Fig. 3
Require: $w$ is the sliding window size
1: for $j = 1$ to $n_i - w$
2:     for $i_2 = 1$ to $w$
3:         if $isFoundInAllImage\left(m^i_{i_2,j}, w\right)$ then
4:             $\mathrm{M} \leftarrow retrieveAllCorrespondedFeature(i, j, i_2, w)$
5:             $\overrightarrow{\psi} \leftarrow interpolate(M)$
6:             $D^i_j \leftarrow addNewEntry\left(D^i_j, \overrightarrow{\psi}\right)$
7:         end if
8:     end
9:     $\mathcal{D}^i \leftarrow addNewEntry\left(\mathcal{D}^i, D^i_j\right)$
10: end
```

scriptors corresponding to the j^{th} window (sub-place), there would be $n_i - w + 1$ dictionary for representing the place when the extraction is completed.

The depiction of Algorithm 1 shows how extraction is performed. Given n_i images of the i^{th} place, and a set of matching index vector $\overrightarrow{m}$ derived from sequential matching in the previous sections, a sliding window of size w is created to extract the PIRF of sub-places. For each sub-place j, all matching index vectors $\overrightarrow{m}$ are processed to find those local features which appear repeatedly in all images of the current window (line 3). All corresponding $w - 1$ features would be retrieved and put into the temporary matrix M if such features were found (line 4). These features are interpolated using averaging to obtain a single representative feature $\overrightarrow{\psi}$. **This feature $\overrightarrow{\psi}^i_{xj}$ is the x^{th} single PIRF of the j^{th} sub-place of the i^{th} place.** Each extracted PIRF is gradually collected into the PIRF-dictionary of the j^{th} sub-place D^i_j (line 6). This extraction process is repeated until the window is slid to the last image of the place. Each sub-place has its own dictionary D^i_j, $j \le n_i - w + 1$. These dictionaries are finally concatenated mutually to form the dictionary of the i^{th} place $\mathcal{D}^i$ (line 9). Given d_j as the number of PIRFs in the dictionary of the j^{th} sub-place, $\mathcal{D}^i$ as the PIRF dictionary of the i^{th} place, and $n^i_{\mathcal{D}}$ as the total number of PIRFs in $\mathcal{D}^i$, the PIRF dictionaries for representing all visited places $\mathbb{D}$ are derived as presented below:

$$D^i_j = \begin{bmatrix} \overrightarrow{\psi}^i_{1,j} \\ \vdots \\ \overrightarrow{\psi}^i_{d_j,j} \end{bmatrix}, \ \mathcal{D}^i = \begin{bmatrix} D^i_1 \\ \vdots \\ D^i_{n_i-w} \end{bmatrix}, \ \mathbb{D} = \begin{bmatrix} \mathcal{D}^1 \\ \vdots \\ \mathcal{D}^{\mathrm{N}} \end{bmatrix} \tag{1}$$

$$n^i_{\mathcal{D}} = \sum_{j=1}^{n_i} d_j \tag{2}$$

Therein, $\mathbb{D}$ is useful to represent all visited areas in the environment. Extraction is incremental because the new area can be simply added to the library. Additionally, it is worth noting that extracted PIRFs must match images only $\left(\sum_{i=1}^{\mathrm{N}} n_i\right) - 1$ times, whereas the spectral clustering (SC) requires $\left(\sum_{i=1}^{\mathrm{N}} n_i\right) \times \left(\left(\sum_{i=1}^{\mathrm{N}} n_i\right) - 1\right)/2$ times to form the affinity matrix. Although incremental spectral clustering (ISC) [15] performs fewer image comparisons than SC, the comparisons are still much more numerous than those using our method.

3.3 Place Recognition

Now that all N individual places, $\mathbb{P} = \{\mathrm{p}_1, \ldots, \mathrm{p}_{\mathrm{N}}\}$, are well represented using a set of corresponding PIRF-dictionaries $\mathbb{D} = \left\{\mathcal{D}^1, \ldots, \mathcal{D}^{\mathrm{N}}\right\}$, we can describe how these PIRF dictionaries are used to recognize places. Majority voting is selected as the recognition framework.

Majority voting (MV) is a popular combination scheme because of its simplicity and its performance on real data. Its performance has been demonstrated experimentally in many studies such as handwriting recognition [40] and personal authentication. We select MV because of its main concept related to the independence of recognizers. Based on theoretical analyses, MV is apparently effective if the recognizers are independent. Considering our problem, we assume that each place is independent. By applying MV to our problem, each place vote for the matched descriptors is found in the testing image. Additionally, MV is suitable for the task of incremental map-building in robotics, as described in [22], because a similarity threshold for image comparison is not needed. The image is assigned to the place with the maximum number of votes.

Consider the problem in which a single omnidirectional image I is to be assigned to one of N possible existing places $(\mathrm{p}_1, \ldots, \mathrm{p}_{\mathrm{N}})$. First image I is extracted and a set of descriptors, $\mathbb{Z} = (\overrightarrow{z}_1, \ldots, \overrightarrow{z}_n)$, is derived, where $\overrightarrow{z}$ is a single image descriptor and n is the number of descriptors. Of N places, each checks if the descriptor $\overrightarrow{z}_k$, $1 \le k \le n$ is similar to any PIRF in its dictionary $\mathcal{D}^i$, $1 \le i \le \mathrm{N}$. The vote is counted and the score is increased by one for every matching: we initialize $S_i \rightarrow 0$ for every i,

$$\begin{aligned} &S_i = S_i + 1 \quad \text{if} \\ &\min_{1 \le j \le n^i_{\mathcal{D}}} |\overrightarrow{z}_k - \overrightarrow{\psi}^i_j| < \tau, \end{aligned} \tag{3}$$

where τ is the similarity threshold for feature matching (we found earlier that $\tau = 0.6$ yields the best performance). The vote from places can be done in parallel, thereby enabling rapid recognition. After voting has been completed, the system recognizes the image I as

$$assign \quad I \rightarrow \mathrm{p}_{\mathrm{argmax}_i(S_i)}$$

with confidence

$$c_{\mathrm{argmax}_i(S_i)} = \frac{S_i}{\sum_{j=1;j\neq i}^{\mathrm{N}}(S_j)} \tag{4}$$

We have now described image classification to an existing class. In the next section, we consider incremental topological mapping by which the input images might come to belong to either an old or a *new* place.

3.4 PIRFs Reduction

In the view of long-term recognition, a main concern for using PIRFs is the amount of PIRF in the current system. One PIRF-dictionary is used as a signature of one place. Therefore, the number of dictionaries depends on the number of visited places. For long-term recognition, the number of places is infinite, which means that the number of PIRF-dictionaries would also be infinite. Two techniques are used to solve this problem: reducing (i) the number of PIRFs or (ii) the number of dictionaries. The first technique is to slow the growth rate of PIRF. However, even though the number of PIRFs grows very slowly, it will finally reach the memory limits. The second technique is used here to delete or forget an unnecessary dictionary. The techniques are simple but efficient.

Reducing PIRFs. Most PIRFs are of distant objects whose appearances are robust against the change of viewpoints along the path. These objects, on the other hand, are also likely to be detected as distant objects in other places as well. For example, Tokyo Tower is visible in many places throughout Tokyo. Seeing the tower does not help in identifying the place. Therefore, the PIRFs which capture the Tokyo Tower are useless and should be deleted. These PIRFs might be treated as "redundant PIRFs". To eliminate these PIRFs, training images can be re-used: some recognized images were retained for the following retest. By this retest, the system knows which PIRFs match to the right object, and which PIRFs do not. Particularly, given $\mathbb{D}_t = \{\mathcal{D}_1, \ldots, \mathcal{D}_{n_t}\}$ as the set of PIRF-dictionary up to time t, and I as an input testing image, the dictionary with the highest matched PIRFs, denoted by $\mathcal{D}_{win}$, wins recognition with confidence (vote quality) c. This recognition result is used to update the scores of all PIRFs in all dictionaries. For every matching between descriptor x of image I and the corresponding PIRF $\overrightarrow{\psi}_i^{win}$ of dictionary $\mathcal{D}_{win}$ where $1 \leq i \leq n_{\mathcal{D}}^{win}$, increase the score α_i^w of the PIRF by 1. In contrast, for every matching between descriptor x of image I and the corresponding PIRF $\overrightarrow{\psi}_i^{\mathbb{D}_t - \{\mathcal{D}_{win}\}}$, decrease the score of the PIRF by 1. A *high-score* PIRF can be inferred as a useful PIRF; it often matches features of a distinctive object in the place, whereas a *low-score* PIRF can be interpreted as a PIRF which either captures confusing objects (an object that is visible from many places) or which captures highly sensitive objects (an object which is visible to only a few camera positions in such places). After a re-test, almost all PIRFs would already be assigned scores. Sorting the PIRFs by their scores, the number of PIRFs can be reduced by rate R (*i.e.*, $R = 0.75$, $R = 0.50$). The reduction drastically shrinks the PIRFs without a marked drop in accuracy. A PIRF with $\alpha = 0$ is simply treated as a redundant PIRF.

Generally, this reduction technique is performed *offline* because it must run a batch retest on previous images to update scores of PIRFs. However, "when to update the scores" is flexible; the robot can wait until it is free to take time thinking of the past and to update the dictionaries. This reduction could be postponed if the robot was busy with some task. This process can also be done *online* by taking advantage of the assumption of physical robots. Actually, scores can be updated every time the system recognizes a new image. What the system must know is whether the recognition result is correct or not. This is solvable by assuming that the robot actually obtains more than two images before making a decision. Therefore, once the system recognizes input image I as place p_w (with corresponding dictionary $\mathcal{D}_w$), it continues recognizing the first and second next images to confirm further that the images really belong to p_w; then it updates the score of PIRFs. Details about this online reduction method are described in the next section of robotic applications.

Figure 4 portrays PIRF scores obtained by running a retest on all 382 training images. Most PIRFs were used at least once. The score separates the good and the bad PIRFs. We later show in Sect. 4 that, even after reducing the size of PIRF by 50%, the recognition rate is still high, which underscores the effectiveness of the reduction.

Forgetting Places. Certainly, long-term recognition will eventually confront the problem of memory overload because the number of places is infinite. For that reason, a

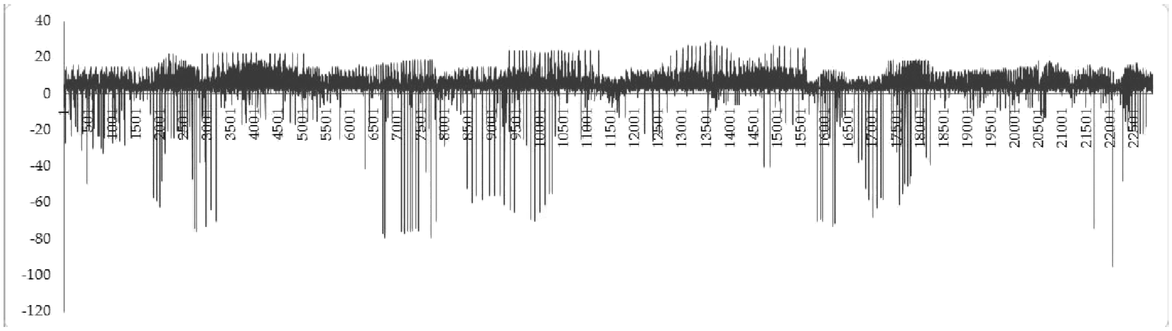
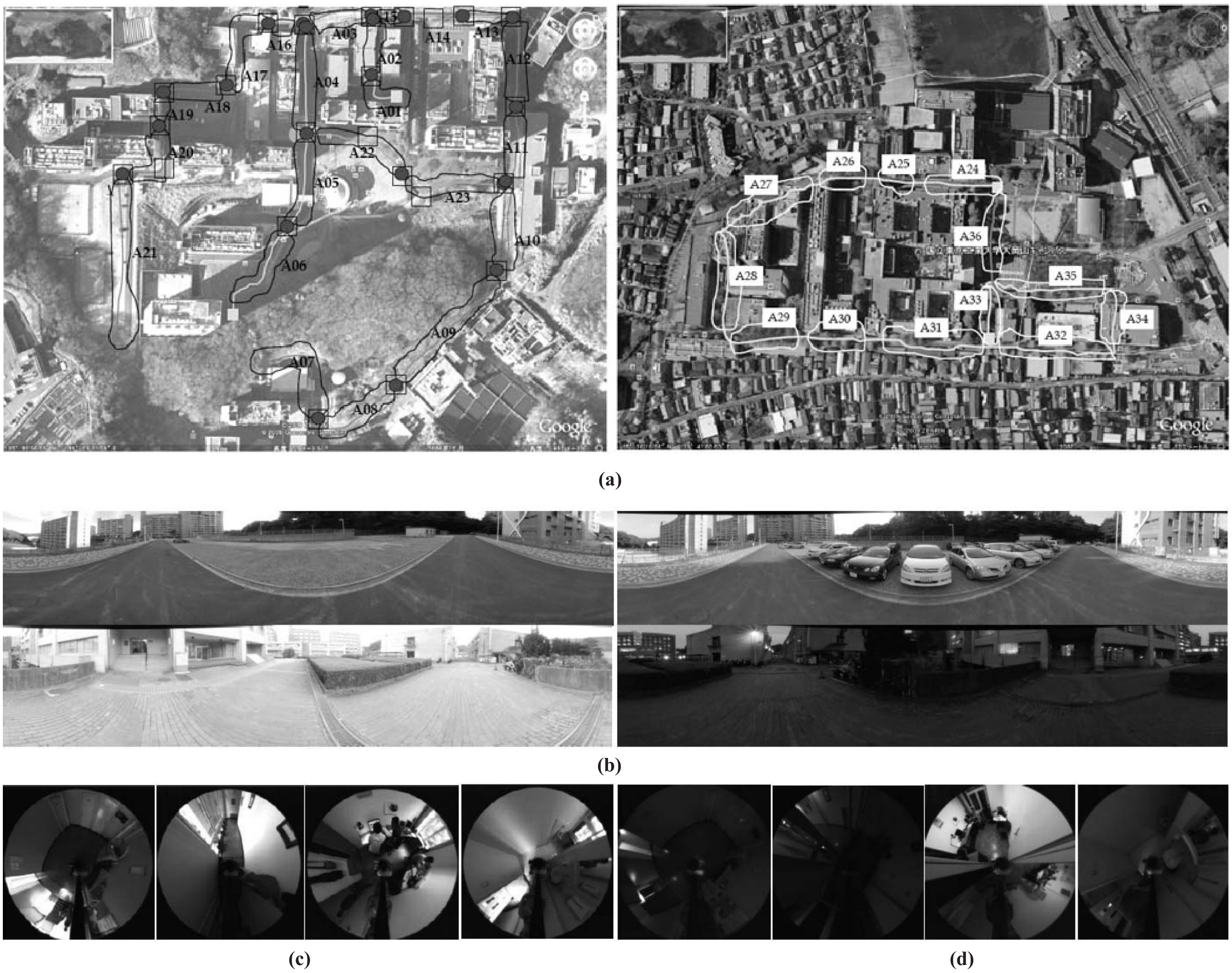


Fig. 4 Updated score of 22901 PIRFs corresponding to 15 places of Suzukakedai Campus. The re-test was done over the same set of 382 images. The x-axis is the frequency score of the PIRFs, the axis is the frequency score of the PIRFs, and the y-axis are the indices of all 22901 PIRFs. We found that the dictionary with a high average score of PIRFs (*i.e.*, PIRFs of index 11414–15629 of the 6^{th} place $\mathcal{D}^6$) is extracted from the isolated place where most of the faraway objects (*i.e.* high building) are not shared by other places.

robot must “forget” some places that are considered to be of no use, or at least temporarily remove such places from the searching space to speed up the localization time. In fact, PIRF-based recognition uses the PIRF-dictionary as the signature of the whole individual place. Once the robot is sure that a place will never be visited again, the corresponding dictionary can be simply deleted, or moved to other memory spaces. Although the environments used for this study are large in scale, they are not so large as to require the place-forgetting procedure. The technique for a larger environment, *i.e*. 100+ places is described here.

4. Application to Robotic Navigation

In robotic topological mapping, determining the number of topological nodes (how to partition image data into classes) is an important concern. Some previous works [21], [31], [36] ignored this problem by partitioning image samples



(a)

(b)

(c)

(d)

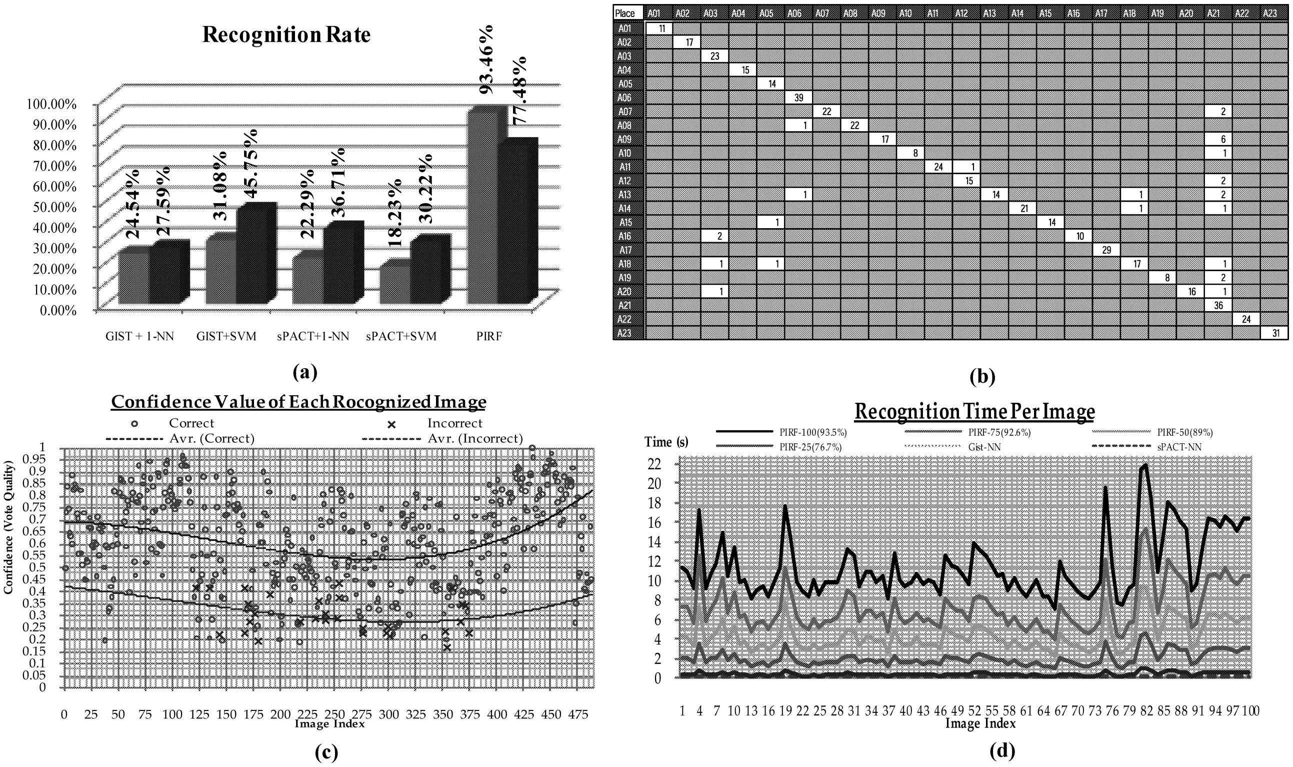
Fig. 5 Map of the outdoor experiment sites. **(a)** Twenty-three places manually segmented by hand from Suzukakedai campus (left). An additional 13 places of Okayama campus were added (right). For Suzukakedai, train data were collected on holidays under clear weather during daytime, whereas test data were collected on weekdays under various conditions (*i.e*. cloudy, sunny, night). For Okayama, train and test data are collected randomly on weekdays under various conditions over 3 months. **(b)** Sample of images from place A21 (top) and A01 (bottom) of Suzukakedai. The training image was collected on a holiday (top-left), whereas the testing image was obtained on weekdays (top-right). Some training images are taken in daytime (bottom-left), while the testing image was obtained in the evening (bottom-right). Both images are unwrapped merely for illustration. **(c)** Example of images of the Ljubljana lab taken by iRobot ATRV-Mini from the COLD database of [33]. From two available sub-datasets, we select the standard path database, which is taken from the *Printer Area, Corridor, A shared office*, and *a bathroom* (shown respectively from left to right). In the study described in this paper, we use two different set of sequences taken under cloudy, sunny, and nighttime conditions, constituting *ca*. 6000/6000 sequential images of 640 × 480 pixels for use in training/testing. **(d)** Samples images taken in nighttime from the same four places as in **(c)**.

manually into classes. Later, Valgren and Lilienthal [15] proposed incremental SC (ISC) so that the algorithm becomes fully incremental. Nevertheless, partitioning based only on appearance might yield too many nodes. For example, the work described in an earlier study [16] has 160 nodes (classes).

In this section, we describe a simple but effective method to use PIRF in topological mapping and localization: the obtained performance is substantially better than that of ISC. Our topological mapping is expected (i) to be fully incremental, and (ii) to output a reasonable number of nodes (matched well to the environment). Using the concept of Spatial Semantic Hierarchy (SSH) described elsewhere in the literature [8], we simply add the junction detection module to the control layers of robot. This module takes an omnidirectional image as input, unwraps it into a panoramic image, and then classifies it as either a *junction* or *non-junction* image. This module would signify the upper recognition system to set the partition boundary if it judged the image as that of a junction. This guarantees that the number of nodes or places in the map matches well to the environment; it depends only on the number of detected junctions. For example, 580 training images of Suzukakedai Campus (see Fig. 5 (a-left)) are segmented into 23 places with 19 junctions. After the partitioned images are obtained, we can perform the recognition as in the previous section.

For this study, we implement a junction detection system resembling that described in an earlier study [14]: color histograms are used instead of Gaussian Mixture Models. The only difference is that our detection is of omnidirectional images. The assumptions resemble those of a prior study [14] where the lowest area of the image is the road; the upper part is the background. The road and off-road area pixels are sampled to create models of 30 × 30 histograms of red (R) and green (G) (because this setup yields the best result according to [14]) for representing the road and the background.

As depicted in Fig. 5 (a-left), junction detection from 580 omnidirectional images is possible (circle). Particularly, the system samples the road and background pixels of each image and then performs binary classification (junction/non-junction). For all 580 images, most junctions were detected correctly with some small error contained (the system detects the junction a few images before or after the correct



Place	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23
A01	11																						
A02		17																					
A03			23																				
A04				15																			
A05					14																		
A06						39																	
A07							22														2		
A08						1		22															
A09									17												6		
A10										8											1		
A11											24	1											
A12												15									2		
A13						1							14					1			2		
A14														21				1			1		
A15					1										14								
A16			2													10							
A17																	29						
A18			1		1													17			1		
A19																			8		2		
A20			1																	16	1		
A21																					36		
A22																						24	
A23																							31

Fig. 6 Recognition Results **(a)** The overall performance of PIRF is shown in comparison with other methods in term of accuracy (Gray->O-okayama, Black->Suzukakedai). **(b-c)** Other results on 489 testing images of Suzukakedai. **(b)** The confusion matrix of recognition results of 489 images of Suzukakedai (Row-> corrected classes, Column-> predicted classes). **(c)** The confidence value of each image recognized by our system. The average lines for correct and wrong recognition are estimated using 3rd order polynomials. This confidence value is similar the quality of vote in the work of [22]. The calculation method is the same. **(d)** The recognition times per image are shown. PIRF-*n* denotes PIRFs reduced to *n*%. The PIRF-50 (Parallel) is the PIRF which has been reduced by 50% and each dictionary vote for the matched PIRF in parallel. A comparison is made with Gist and sPACT.

one). Precisely, four junctions were missed out of all 25 junctions. The image is first converted to a panoramic image and is segmented by the system. The segmentation would then be cropped at the part believed to contain all road vanishing points (we manually set the cropping size to 0.2–0.5 h, where h is the image height). Then, this cropped image is filtered using a Gaussian filter with sigma = 1. The resulting image is averaged by each column to derive the vector which would be smoothed using a pseudo-Gaussian. The number of valid paths is determined by the number of peaks above the mean value. Because of limited space, we do not describe about the junction method in detail.

Now that the junctions have been detected and all images have been partitioned into places, localization is performed. Robot localization might differ slightly from scene recognition; the recognition results of some images in the current place are obtainable before making a decision. Therefore, one step of the robot might contain n_t images, where t is a time step. For each step, the robot recognizes all n_t images and then summarizes the votes of the nearest place p^* with reliability score r^*. At this point, all n_t images would have been recognized. The PIRFs would have been extracted from these images and would have been collected to the new dictionary $\mathcal{D}^{new}$. If the score r^* is greater than threshold θ, then the current place recognized as p^* and $\mathcal{D}^{new}$ is neglected. Otherwise, the place is a new previously unseen place. A new dictionary $\mathcal{D}^{new}$ would be augmented to a set of dictionaries, $\mathbb{D}_{t+1} = \mathbb{D}_t \cup \{\mathcal{D}^{new}\} = \left\{\mathcal{D}^1, \ldots, \mathcal{D}^N, \mathcal{D}^{new}\right\}$.

To examine the process in more detail, both the average value of confidence and the maximum rate of recognition for the nearest place are considered to calculate the reliability score r. Given $\mathbb{I}_t = \{I_1, \ldots, I_{n_t}\}$ as the sequentially observed image of the current place ($n_t \geq 2$), with p_i and c_i respectively signifying the assigned place and confidence. The binary valued function as

$$\Delta_{ki} = \begin{cases} 1 & \text{if } p_i = k \\ 0 & \text{Otherwise} \end{cases} \tag{5}$$

and p^* as the nearest place class to which most input images have been assigned, where the following hold

$$p^* = p_j \quad \text{if} \\ \sum_{i=1}^{n_t} \Delta_{ij} = \max_{1 \leq k \leq N} \sum_{i=1}^{n_t} \Delta_{ki}. \tag{6}$$

The set of images $\mathbb{I}$ is recognized as place $p^* = p_j$ if

$$r^* = r_j = \frac{1}{n_t}\left[\omega_1 \cdot \left(\sum_{i=1}^{m_t} \Delta_{ij}\right) + \omega_2 \left(\sum_{i=1}^{m_t} \Delta_{ij} c_i\right)\right] < \theta \tag{7}$$

where θ is the threshold set by the user. For this study, we use $\theta = 0.6$, $\omega_1 = 0.6$, and $\omega_2 = 0.4$ (the importance lays on the number of correct votes. Results show that recognizing new places as existing classes usually elicits low scores, whereas recognizing old places to the corresponding class gains a much higher score (see Fig. 6 (c)). However, PIRF-Nav requires at least two places (nodes) in the map as an initialization before starting the incremental process. While executing, the system obtains new input images and makes a decision for each input image.

As described in previous section, the PIRF reduction can be done in an online manner. Instead of running a batch retest with a great number of re-used images, we incrementally update the score for every recognition step using the real testing images.

5. Experiments and Results

Four main experiments are done to prove the advantages of PIRF. Experiments 1, 2, and 3 examine scene recognition in both indoor and outdoor scenes. These experiments show that PIRF offers a markedly better rate of accuracy than other features for the task of highly dynamic outdoor scenes, while retaining good result for indoor. Experiment 4 is to show that the PIRF-based navigation system outperforms ISC in terms of time and accuracy.

Three image databases were used for this study: *Suzukakedai Campus*, *O-okayama Campus* and *COLD*. Cognitive Systems for the Cognitive Assistants Localization Database (COLD) dataset [33] was captured in a four-room office environment, including *a printer area, corridor, two-person office*, and *a bathroom*. Images were taken by a robot. The purpose of this dataset is to recognize which room the robot is in based on a single image. First regarding the outdoor images datasets (Suzukakedai Campus and O-okayama Campus), we collected them by setting a tripod with height *ca.* 1.7 m. mounted with a camera (60D, DSLR; Nikon Corp.) with an omnidirectional lens. We walk along the road on campus while capturing omnidirectional images every few meters. The camera positions along the road are various (*i.e.*, the position must be moved to the footpath when the car is passing). The images are taken without concern about pedestrians or cars running past by. Some images contain a big blurred object, which actually is a running car. All images' original resolutions were 3872 × 2592, but they were scaled down to 640 × 428 for use in all experiments. For Suzukakedai Campus, most training data were collected on *holidays* under clear weather, whereas the testing data are collected on *weekdays* under various weather conditions, resulting in 580 images for training, and 489 images for testing. All images were collected according to all three routes portrayed in Fig. 5 (a-left). For O-okayama Campus, we collected more images from places A24-A36 in respect to the path portrayed in Fig. 5 (a-right). For this campus, people crowded in the images taken on both holidays and weekdays, so all data were taken on weekdays at various times and weather conditions. Data were collected during 3 months, resulting in 450 images in all for training, and 493 images for testing. Figure 5 (b) portrays differences between training images and testing images. All experiments were written and run using software (Matlab 7.6.0.; The MathWorks, Inc.).

5.1 Experiment 1: Recognizing Outdoor Scenes

This experiment is further divided into two sub-experiments conducted respectively at Suzukakedai campus and Okayama campus. At Suzukakedai, 580 omnidirectional images were input to the system for training. The testing is done over 489 testing images. The training images are labeled by hand according to the junction detection result obtained in experiment 4, resulting in 23 separate places (A01–A23). We do this because, in the discussion related to experiment 4, we can directly use the accuracy of this experiment as the accuracy of PIRF for comparison to the ISC method in experiment 4 (Sect. 5.4). The difficulty of these datasets is the great difference between training and testing data. Changes occurring between images taken on holidays and weekdays are considerable, as shown in the sample image of Fig. 5 (b).

For O-okayama, the dataset is a bit different from Suzukakedai. O-okayama images were not collected on different holidays and weekdays because this main campus is always crowded on both holidays and weekdays. Data were collected during 3 months to attempt recognition despite changes occurring over a long period of time.

Two baselines were used for comparison. The first one (i) is the 80-D Gist vectors used in the work of Torralba et al. [21]. With six orientations of steerable pyramid and four scales applied to the monochrome image, 580 Gist vectors were derived from 580 training images. However, we do not use the HMM as in [21] because the transition matrix of labeled sequence data is not available. Therefore, we try to use First Nearest Neighbor (1-NN) and Support Vector Machines (SVM) as the classification framework for Gist. We also tried 3-NN and 5-NN, but the results are mostly equivalent to those of 1-NN. For the second baseline (ii), the spatial Principal component Analysis on Census Transform (sPACT) proposed by [36] is our choice because of its recently highest result for indoor scene recognition over the IDOL database of [31]. The training images are first converted by the census transformed (CT) image; then the CT histograms are created. Principle Component Analysis (PCA) is then performed on the CT histograms to extract the most important components among the distribution of CT histograms. In this study, we also apply the level 2 spatial pyramid, as done in [36]. As hinted in [36], choosing the right classifier for a specific application is important. Consequently, the classifiers used with sPACT are both NN and SVM, in the same way as that done for the first baseline.

Results are presented in Fig. 6 (a). Actually, PIRF-based recognition yields about a two times higher rate than the others both for Suzukakedai and O-okayama. It is noteworthy that the recognition rate is considerably lower, at only 77.48%. We suspect that this occurs because Okayama campus is a main campus crowded with many people. Most parts of campus are not wide open compared to the Suzukakedai campus. The campus contains crowded buildings so that the problem of perceptual aliasing occurs. The Okayama campus contains many artificial structures that make it similar to indoor areas with highly dynamical changes. Suzukakedai contains more natural distant objects than Okayama, thus having a higher rate.

We also report the confusion matrix of Suzukakedai in Fig. 6 (b): errors are, in general, not distributed uniformly. Taking a look into the place A21 (which confuses many places), for instance, we found that the place is a large open-wide area (sample of image from A21 is portrayed in Fig. 5 (b-up) where many distant objects are shared with other places). Although A21 shared some distant objects with many places, place recognition is still efficient because votes on other objects are useful for determination.

We believe that some reasons make the PIRF-based recognizer outperform others in this experiment. Distant objects, which are robust to positional changes, usually appear smaller than nearby objects. Consequently, global features that capture a whole scene, such as Gist, include many unstable objects, *e.g.*, cars, doors, a gate, people. The sPACT provides a lower rate of recognition than our PIRF because of its basic nature of feature extraction; sPACT converts the whole image into the Census Transformed. Although its performance is recently considered the highest for the IDOL database, its accuracy is lower when tested on our highly dynamic outdoor scenes. Although sPACT is a local descriptor with greater descriptive power than Gist features, sPACT still includes many dynamic objects of the scenes. Being sensitive to changes in camera positions, nearby objects can strongly affect the recognition system.

Distinctive from the others, PIRF assigns emphasis to distant objects while neglecting most nearby objects. The underlying concept of PIRF-recognition differs from those methods that perform segmentation (*i.e.*, normalized cut) before feature extraction. In fact, a PIRF can be extracted without going offline for image segmentation. Unlike the BoW approach, PIRF does not quantize the descriptors; it can therefore preserve the distinctiveness of original local descriptors. Consequently, PIRF can mostly overcome the problem of highly dynamic changes because almost all unstable closed objects are ignored (see Fig. 10 for sample-matched PIRF in testing images).

The confidence values in Fig. 6 (c) underscore the quality of the recognition provided by PIRF. Almost correct recognitions display high confidence values, which can be interpreted as the quality of vote in the same sense of [22]. This fact proves that PIRF is sufficiently discriminative for place identification; the right dictionary only matches to the right place. Consider Fig. 5 (b-top), for example, with two different scenes taken on a holiday and weekday. While training on holiday images (left), PIRF captures the distant building because its appearance is robust to position changes. Therefore, although the testing image might be changed dynamically and might share some distant objects with other places, the number of votes on distant buildings would be adequately higher than other votes. However, it should be noted that, during image index 262–301 (see Fig. 6 (c)), the confidence values are quite low because place A09 is mostly covered by trees along both sides of the path. These trees block the view of distant buildings. Therefore, distant objects are insufficient for justification. Nevertheless, the recognition rate remains good because some nearer objects, which are apparently stable for sub-places, have been used instead.

With respect to the learning time used for model training and feature extraction, PIRF is faster than sPACT or Gist. For every image in the Suzukakedai Campus, the average times for creating the CT histogram, Gist and PIRF are, respectively, 29.2931 s, 4.8219 s, and 3.2312 s. Based on this result, PIRF and Gist are suitable for learning in real-time applications. A comparison of two images can be done quickly during the robot's exploration. In terms of recognition time (per image), PIRF is clearly slower than Gist or sPACT because each encodes an image into only one feature vector. In fact, PIRF trades off the recognition time for better accuracy. However, the recognition time of PIRF-based method can be reduced further using the reduction technique described in Sect. 2.4 to reduce the number of PIRFs and slow the PIRF growth rate. Figure 6 (d) presents the accuracy values obtained for different number of PIRFs. Even with 50% reduction of PIRF, the accuracy remains higher than the other baselines. It is particularly interesting that parallel votes (each dictionary votes simultaneously) can reduce the time to less than a second per image (Lowest Thick Line in Fig. 6 (d)). Although that reduced time is still longer than that of Gist, it might be acceptable for robotic navigation, for which the image capture rate must correspond to the robot's motions. Experiment 4 will show that PIRF-Nav executes more quickly than ISC, although its recognition time is longer than that of either sPACT or Gist.

5.2 Experiment 1: Recognizing Outdoor Scenes

In the second experiment, the PIRF are tested on COsy Localization Database (COLD) of [33]. The COLD database includes data collected by three robots. In this work, we select the data collected from *Ljubljana* laboratory (we designate the experiment using this database as "Ljubljana" hereinafter). We select two sequences of each weather condition: cloudy 1 & 2, sunny 1 & 2, and night 1 & 2, in all six sequences (*ca.* 2000 images for each sequence). Training is done on one sequence and the testing is done on the rest. We repeat the experiments six times (all combinations) and obtain averaged results as portrayed in Fig. 7. Although we clearly claimed that PIRF is especially suitable for outdoor scenes, it is also very common for any long-term system to recognize both outdoor and indoor scenes. To prove that PIRF can also be used indoors efficiently, we conduct this experiment. Furthermore, this experiment proves that PIRF is applicable to images collected by a real robot at various times and in weather conditions (Fig. 5 (c-d) portrays the sample images of COLD taken in sunny and at night. The baselines used in this experiment are the same as those used in experiment 1: sPACT and Gist, and the referred result of [33].

Although Gist and sPACT yield the highest accuracy for this indoor database, PIRF also works well. Especially during daytime (trained or tested with either sunny or cloudy), PIRF offers a high rate of about 93%–94%. In daytime, the scene is clear and a sufficient number of extracted SIFTs are used for PIRF generation. With more PIRFs, the vote quality is high: it can recognize either sunny or cloudy scenes correctly. When testing at night, although the number of PIRFs is sufficient for recognition, the number of SIFT extracted from a testing image is insufficient for representing an image. It is also true that darkness can reduce the number of SIFTs. Consequently, the quality of votes for the nearest place is low. It is also noteworthy that the unbalanced number of sample images does not affect the performance of PIRF-based systems. In this experiment, a corridor was traversed many times by the robot, gathering *ca.* 1500 sequential images. The PIRF-dictionary of corridor is also the biggest. However, the confusion matrix in Fig. 8 shows that the different dictionary size does not engender biased recognition.

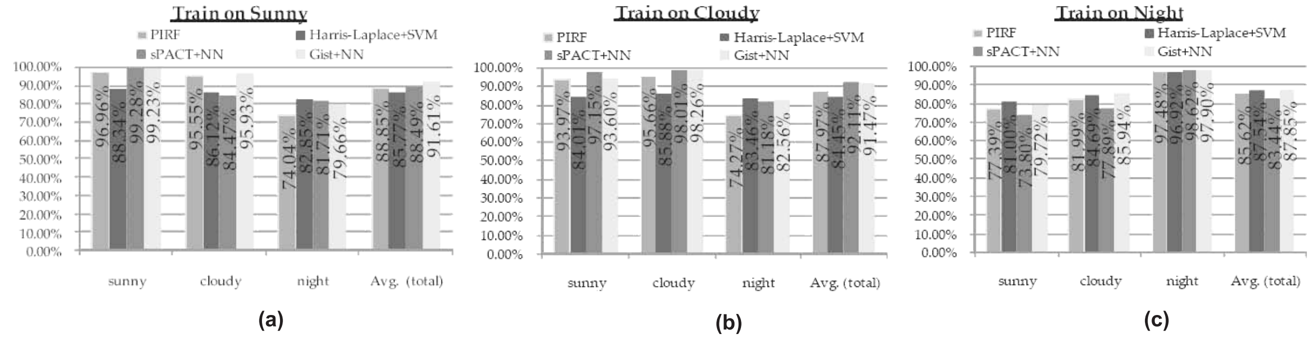
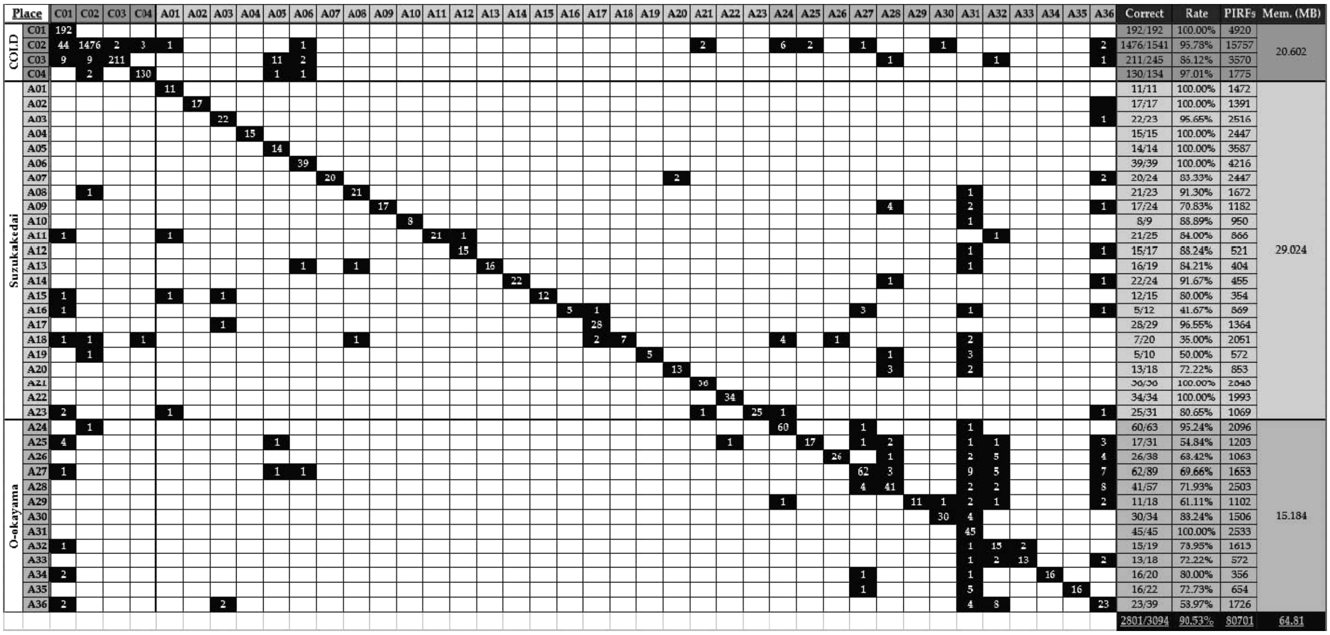


Fig. 7 Averaged recognition results with COLD-Ljublajana standard sequences. Training is done on one sequence and testing on the other sequences. **(a)** Train on sunny. **(b)** Train on cloudy. **(c)** Train on night. The comparisons are done to sPACT of [36] using NN as classifier, to Gist of [21] using NN as classifier, and to results of Harris-Laplace with SVM [33].



Place		C01	C02	C03	C04	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34	A35	A36	Correct	Rate	PIRFs	Mem. (MB)
COLD	C01	192																																								192/192	100.00%	4920	20.602
	C02	44	1476	2	3	1					1															2			6	2		1			1						2	1476/1541	95.78%	15757	
	C03	9	9	211						11	2																						1				1				1	211/245	86.12%	3570	
	C04		2		130					1	1																															130/134	97.01%	1775	
Suzukakedai	A01					11																																				11/11	100.00%	1472	29.024
	A02						17																																			17/17	100.00%	1391	
	A03							22																																	1	22/23	95.65%	2516	
	A04								15																																	15/15	100.00%	2447	
	A05									14																																14/14	100.00%	3587	
	A06										39																															39/39	100.00%	4216	
	A07											20													2																2	20/24	83.33%	2447	
	A08		1										21																							1						21/23	91.30%	1672	
	A09													17																			4			2					1	17/24	70.83%	1182	
	A10														8																					1						8/9	88.89%	950	
	A11	1				1										21	1																				1					21/25	84.00%	866	
	A12																15																			1					1	15/17	88.24%	521	
	A13										1		1					16																		1						16/19	84.21%	404	
	A14																		22														1								1	22/24	91.67%	455	
	A15	1				1		1												12																						12/15	80.00%	354	
	A16	1																			5	1										3				1					1	5/12	41.67%	869	
	A17							1														28																				28/29	96.55%	1364	
	A18	1	1		1								1									2	7						4		1					2						7/20	35.00%	2051	
	A19		1																					5									1			3						5/10	50.00%	572	
	A20																								13								3			2						13/18	72.22%	853	
	A21																									36																36/36	100.00%	2048	
	A22																										34															34/34	100.00%	1993	
	A23	2				1																				1		25	1												1	25/31	80.65%	1069	
O-okayama	A24		1																										60			1				1						60/63	95.24%	2096	15.184
	A25	4								1																	1			17		1	2			1	1				3	17/31	54.84%	1203	
	A26																														26		1			2	5				4	26/38	68.42%	1063	
	A27	1								1	1																					62	3			9	5				7	62/89	69.66%	1653	
	A28																															4	41			2	2				8	41/57	71.93%	2503	
	A29																												1					11	1	2	1				2	11/18	61.11%	1102	
	A30																																		30	4						30/34	88.24%	1506	
	A31																																			45						45/45	100.00%	2533	
	A32	1																																		1	15	2				15/19	78.95%	1613	
	A33																																			1	2	13			2	13/18	72.22%	572	
	A34	2																														1				1			16			16/20	80.00%	356	
	A35																															1				5				16		16/22	72.73%	654	
	A36	2						2																												4	8				23	23/39	58.97%	1726	
																																										2801/3094	90.53%	80701	64.81

Fig. 8 Results of combined sites. The matrix shows confusion among 40 dictionaries of 40 places (Row-> corrected classes, Column-> predicted classes). The number of PIRFs for each place and the memory required are shown.

5.3 Experiment 3: Combined Sites

In the experiment, we combine all experiment sites together: Suzukakedai, Okayama, and Ljubljana sunny 1 & cloudy 1 sequences. Testing images from outdoor scenes are the same as those used in experiment 1, although the testing images for Ljubljana are those of cloudy 1 (trained by sunny 1). This experiment shows that PIRF-dictionaries are sufficiently distinctive even for recognizing a larger environment. The efficiency is unaffected by the imbalance of training samples in places with different sizes (*i.e.*, the corridor is longer than the bathroom). The results are shown as a confusion matrix in Fig. 8.

Torralba et al. [21] navigated the robot indoors and outdoors. We do the same by combining the PIRF-dictionaries of Suzukakedai, Okayama, and indoor Ljubljana. The testing images are the same set used previously. The results are presented in Fig. 8. Results show that the accuracies of PIRF are approximately equal—2009/2112 (95.12%), 423/489 (86.50%), and 375/493 (76.06%)—respectively, for Ljubljana, Suzukakedai and O-okayama. The PIRF-dictionary is not affected by the imbalanced image samples in each place.

Based on the obtained results, we conclude that two main factors affect the PIRF's efficiency. **(i)** The characteristics of the place. Places with numerous objects blocking the distant view of the camera inject bad PIRFs into the dictionary. With only a few distant objects, the PIRF must capture some nearby objects instead. In addition, in cloudy weather, sometimes a distant view in an image is too bright (this is the problem of photography in which the illumination condition the space between the distant view and the camera position is too different). In this case, only a very few SIFTs would be extracted from distant objects. **(ii)** The size of the place. A few image samples can cause failure of PIRF extraction in the sense that only a few wide-baseline features were found. For example, A16 and A19 obtained a low rate of accuracy because they are much smaller than other places, whereas A18 obtains low accuracy because its high slope blocks most of the distant views. It must be clarified that the imbalance of sample images of places does not affect the recognition as long as the PIRFs in the dictionary are sufficiently distinctive; this depends directly on the characteristics of the place. For example, 1104 PIRFs are sufficient for representing A01. The numerous PIRFs of C02 (15757 PIRFs) cannot confuse A01, although only 2533 PIRFs of A31 confuses many places. Examining place A31 (which confuses many places) for instance, we found that the places are also open-wide areas where many distant objects are shared with other places. Several buildings are visible in this place. It is important to note that PIRF suits a wide-open area: wide-open areas (*i.e.* A31, A21) themselves obtain a very high rate of recognition (100%), but they can also confuse other areas. This might be resolved simply by re-examining the confusion matrix and deleting some PIRFs that are confusing. Nevertheless, overall, the results show that the PIRF-dictionary is sufficiently distinctive to offer a better recognition rate than other features.

5.4 Experiment 4: Inc. Topological Mapping

In this experiment, we show that PIRF is useful to solve appearance-based topological mapping in an incremental manner like that of ISC, but with less computation time. The baseline used in this experiment is the incremental spectral clustering (ISC) of [15], [16]. We let both PIRF-based navigation (PIRF-Nav) and ISC create the map incrementally. The loops have been closed with neither false negatives nor false positives, although the junction detection missed 4 junctions from among all 25 junctions (16.0% false negative). A comparison of time between ISC and PIRF-Nav is presented in Fig. 9.

In terms of mapping time, PIRF-Nav builds the map in *ca.* 30% of the usual time and tends to do so faster in larger environments (Fig. 9 (top-row)). The ISC uses much more time because of its necessary affinity matrix generation. Precisely, ISC requires 167910 comparisons (46888.28 s), ISC requires 40757 comparisons (11996.80 s), and PIRF requires 579 comparisons (3723.23 s). In terms of recognition time, PIRF-Nav also recognizes a single-image rapidly (Fig. 9 (bottom)). The ISC clusters the map into 159 nodes. Therefore, the minimum number of comparisons necessary for each recognition is 159. We further reduce the number of PIRFs to reduce the localization time. We set the reduction rate to 25%, 50%, and 75% to reduce the PIRFs, and classify the testing images again. The result presented in Fig. 9 (bottom) respectively shows that the 75% and 50% reduced PIRFs still yield the same accuracy despite reducing the localization time by *ca.* 25% and 75%.

Regarding accuracy, we implement the standard spectral clustering (SC) as our baseline instead of ISC because SC yields a better rate of classification if k is appropriated. Therefore, we ran SC for many different k and found that $k = 43$ offers the highest accuracy. To classify the images by SC without its associated position data, we simply label 489 testing images with respect to Fig. 5 (a-left). For example, we consider that SC clusters training images no. 1–10 as cluster 1, and images no. 11–25 as cluster 2. Consequently, the testing images taken for area A01 according to Fig. 5 (a-left) are expected to match the images in either cluster 1 or 2. As such, classification of the image from A01 by SC would be considered correct if the nearest cluster is 1 or 2. As a result, SC offers 40.29%, while PIRFs offers 93.46%, two times higher than SC.

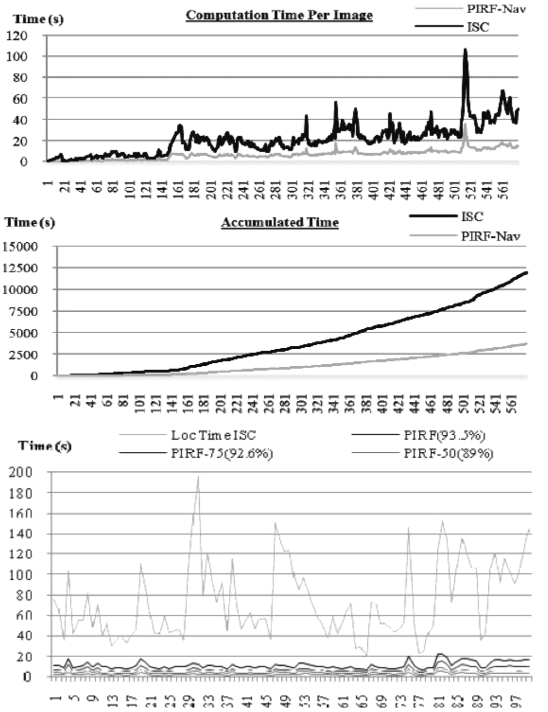


Fig. 9 Comparison of the computation time (top) and accumulated computation time excluding feature extraction for ISC and PIRF-Nav (middle). (Bottom) Computation time for each recognition and the corresponding accuracy. Times are shown for the first 100 testing images. PIRF-75, -50, and -25 have 25%, 50% and 75% respective size reduction. (Programmed by Matlab)

Both SC and ISC represent places with a set of reference images. In a highly dynamic environment, training images taken on holidays appear very different from testing images taken on weekdays. This can cause failure of the recognition. For example, A03 depicts the main road at the entrance of the campus; A21 shows the parking areas. These two places look very different between peak and off-peak times. A comparison between SC and PIRF might imply that several raw SIFTs cannot recognize the highly dynamic scenes. By SC, a set of reference images are retained for matching in the recognition process. The matching is done by local feature matching. If a major portion of an image is changed, then major SIFTs of training images might be unable to match those SIFTs in the testing image.

6. Discussion and Conclusions

The results obtained by our experiments show that PIRF recognizes large sets of both indoor and outdoor images without the help of supervised learning tools such as SVM or HMM. Precisely capturing the points of interest from distant objects, the number of local descriptors can be markedly reduced while preserving their discriminative power.

Regarding accuracy, PIRF clearly outperforms other features in outdoor scenes; it does well even in an indoor environment where distant objects are not ubiquitous. Instead of natural distant objects, PIRF captures nearby objects with a stable appearance. Figure 10 presents samples of testing images with PIRFs matched to the correct dictionaries for both indoors and outdoors. In Fig. 10 (Bottom) PIRF captures the decorations on the wall instead of distant objects.

One concern related to long-term scene recognition is that the number of samples for each place is unbalanced because some places take less time to walk through, but others can take much more time (*i.e.*, the "corridor" dictionary from Ljubljana contains more than 10,000+ PIRFs, whereas the dictionary of "bathroom" contains only *ca.* 2000 PIRFs). Regarding this concern, thanks to the discriminative power of SIFT, all 40 combined dictionaries of PIRFs are sufficiently distinctive to support recognition.

Another remarkable advantage of PIRF is the reduction rate of memory. Because PIRFs are sufficient to represent the place, the reference images are no longer needed. Most previous approaches work with a database of refer-

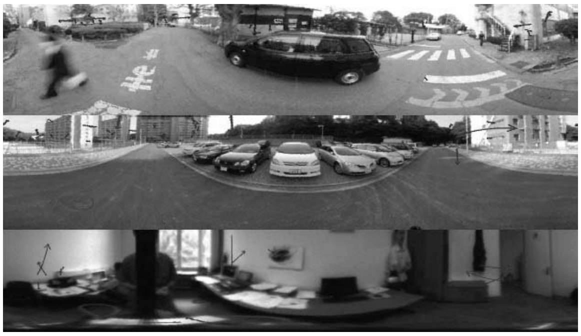


Fig. 10 The sample images described by PIRF. The images are taken from A28, A21, and C03 respectively. For outdoor scenes, the PIRF tries to describe the distant object such as building. For indoor, it captures the most stable objects such as walls.

ence images [10], [19] for which the size depends heavily on the area size. As depicted in Fig. 8, the memory necessary to store all 3091 images (for combined sites) is *ca.* 991 MB, whereas the memory required for storing PIRFs is 57.65 MB (71669 PIRFs, no reduction). Therefore, using PIRF reduces the necessary memory size by *ca.* 94%.

The junction detection module, which is used to partition data into classes instead of clustering, is not the main emphasis of this study. This problem might be regarded as a problem of robot perception. A robot who cannot detect the junction would be like a man who forgets to notice an intersection. Such a junction can be treated simply as a normal straight path. Moreover, without a junction detection system, PIRF can be used simply with preliminary partitioned data like those described in earlier studies [31], [32], [36].

A profound effect of using PIRF is the utilization of stable distant object information. However, PIRF has some limitations and limited future research directions for improvement. One disadvantage of the current PIRF implementation is that (i) it strongly relies on the efficiency of the local descriptors. SIFT is a highly discriminative local descriptor. Therefore, the extracted PIRF can capture distinctive features from objects precisely. On the other hand, this disadvantage makes PIRF flexible for use with other local descriptors, *i.e.* speeded up robust feature (SURF). Second, (ii) PIRF requires input images as the sequences. Although we have claimed that PIRF can solve the kidnapped robot problem (appearance-based localization), it requires that the length of image sequence be sufficient for PIRF extraction (*i.e.* three images). In other words, PIRF is currently limited to the recognition problem; its descriptive power is too great to be used in the problem of categorization or understanding. Third, in this paper, we use PIRF in a simple manner to recognize scenes. (iii) Collecting numerous PIRFs from many places might finally produce a problem of duplicated features. In addition to our PIRF reduction, vector quantization might be another good choice. Because PIRF is a distinctive feature in a highly dynamic environment, bag-of-PIRF might be a good solution for highly dynamic environments. Finally, although the time in PIRF extraction is faster than that of either Gist or sPACT, its recognition time is slower (although it is faster than ISC). In this study, one place required about 500–1000 PIRFs for representation. Room exists for improvement here; one might be able to compress these PIRFs further to speed up the recognition time.

The PIRF features, despite their simple implementation, can achieve promising localization performance, especially in terms of computation time. Comparing the current ISC method [16], PIRF is useful to build the topological map incrementally in considerably less time. Although the current PIRF still requires more than a single image for PIRF extraction, the current result appears promising for future improvement. The localization time (single-image classification) is also shortened considerably because the dictionaries are sufficient for place recognition instead of databases of reference images. Importantly, we do not claim that PIRF-Nav is the most suitable navigation approach for PIRF. We merely describe a simple navigation approach to demonstrate that PIRF is useful for the robotic development community. The standard local descriptors used by the BoW approach [23], [29] do not perform well in highly dynamic scenes. We believe that one might create more efficient robot navigation by considering PIRF, *i.e.* BoW created from three feature spaces, one of which is PIRF.

Acknowledgments

This research was supported by the Industrial Technology Research Grant Program in 2004 from the New Energy and Industrial Technology Development Organization (NEDO) of Japan.

References

[1] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," IEEE Robot. Autom. Mag., vol.13, no.2, pp.99–110, 2006.

[2] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM): Part II," IEEE Robot. Autom. Mag., vol.13, no.3, pp.108–117, 2006.

[3] O.M. Mozos, C. Stachniss, and W. Burgard, "Supervised learning of places from range data using adaboost," Proc. IEEE Int'l Conf. Robotics and Automation, 2005.

[4] Y. Abe, M. Shikano, T. Fukuda, and F. Arai, "Vision based navigation system for autonomous mobile robot with global matching," Proc. IEEE Int'l Conf. Robotics and Automation, pp.1299–1304, 1999.

[5] S. Thrun, "Finding landmarks for mobile robot navigation," Proc. IEEE Int'l Conf. Robotics and Automation, pp.958–963, 1998.

[6] S. Maeyama, A. Ohya, and S. Yuta, "Long distance outdoor navigation of an autonomous mobile robot by playback of perceived route map," Proc. Int'l Symp. Experimental Robotics, pp.185–194, 1997.

[7] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.

[8] B. Kuipers, "The spatial semantic hierarchy," Artif. Intell., vol.119, no.1-2, pp.191–233, 2000.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2006.

[10] T. Goedeme, M. Nuttin, T. Tuytelaars, and L.V. Gool, "Omnidirectional vision based topological navigation," Int. J. Comput. Vis., vol.74, no.3, pp.219–236, 2007.

[11] J. Kosecka and F. Li, "Vision based topological Markov localization," Proc. IEEE Int'l Conf. Robotics and Automation, 2004.

[12] L. Ledwich and S. Williams, "Reduced SIFT features for image retrieval and indoor localisation," Proc. Aust. Conf. Robotics and Automation, 2004.

[13] H. Bay, T. Tuytelaars, and L.V. Gool, "SURF: Speeded up robust features," Proc. European Conf. Computer Vision, 2006.

[14] C. Tan, T. Hong, T. Chang, and M. Shneier, "Color model-based real-time learning for road following," Proc. Int'l IEEE Intelligent Transportation Systems Conf., 2006.

[15] C. Valgren, T. Duckett, and A. Lilienthal, "Incremental spectral clustering and its application to topological mapping," Proc. IEEE Int'l Conf. Robotics and Automation, 2007.

[16] C. Valgren and A. Lilienthal, "Incremental spectral clustering and seasons: Appearance-based localization in outdoor environments," Proc. IEEE Int'l Conf. Robotics and Automation, 2008.

[17] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.888–905, 2000.

[18] L. Renniger and J. Malik, "When is scene identification just texture recognition?," Vision Research, vol.44, pp.2301–2311, 2004.

[19] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," Proc. IEEE Int'l Conf. Robotics and Automation, pp.1023–1029, 2000.

[20] A. Oliva and A. Torralba, "Modeling the shape of scene: A holistic representation of the spatial envelope," Int. J. Comput. Vis., vol.42, no.3, pp.145–175, 2001.

[21] A. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin, "Context-based vision system for place and object recognition," Proc. IEEE Int'l Conf. Computer Vision, pp.1023–1029, 2003.

[22] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," Proc. IEEE Int'l Conf. Robotics and Automation, pp.3921–3926, 2007.

[23] A. Angeli, D. Filliat, S. Doncieux, and J. Way, "Fast and incremental method for loop-closure detection using bags of visual words," IEEE Trans. Robotics, vol.24, no.5, pp.1027–1037, 2008.

[24] A.C. Murillo, C. Sagues, J.J. Guerero, T. Goedeme, T. Tuytelaars, and L.V. Gool, "From omnidirectional images to hierarchical localization," Robotics and Autonomous Systems, vol.55, no.5, pp.372–382, 2007.

[25] E. Royer, M. Lhuillier, M. Dhome, and J.M. Lavest, "Monocular vision for mobile robot localization and autonomous navigation," Int. J. Comput. Vis., vol.74, no.3, pp.237–260, 2007.

[26] L.A. Clemente, A.J. Davison, I.D. Reid, J. Neira, and J.D. Tardos, "Mapping large loops with a single hand-held camera," Proc. Robotics: Sciences and Systems, 2007.

[27] H. Andreasson, T. Deckett, and A.J. Lilienthal, "A minimalistic approach to appearance-based visual SLAM," IEEE Trans. Robotics, vol.24, no.5, pp.991–1001, 2008.

[28] P. Newman, D. Cole, and K. Ho, "Outdoor SLAM using visual appearance and laser ranging," Proc. IEEE Int'l Conf. Robotics and Automation, pp.1180–1187, 2006.

[29] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," Int. J. Robot. Res., vol.27, pp.647–665, 2008.

[30] A. Tapus and R. Siegwart, "Incremental robot mapping with fingerprints of places," Proc. IEEE Int'l Conf. Intelligent Robots and Systems, 2005.

[31] A. Pronobis, B. Caputo, P. Jensfelt, and H.I. Christensen, "A discriminative approach to robust visual place recognition," Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems, 2006.

[32] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "Incremental learning for place recognition in dynamic environments," Proc. IEEE Int'l Conf. Intelligent Robots and Systems, 2007.

[33] M.M. Ullah, A. Pronobis, B. Caputo, J. Luo, P. Jensfelt, and H.I. Christensen, "Towards robust place recognition for robot localization," Proc. IEEE Int'l Conf. Robotics and Automation, pp.530–537, 2008.

[34] J. Civera, A.J. Davison, and J.M.M. Montiel, "Inverse depth parameterization for monocular SLAM," IEEE Trans. Robotics, vol.24, no.5, pp.932–945, 2008.

[35] X. Ren and J. Malik, "Learning a classification models for segmentation," Proc. IEEE Int'l Conf. Computer Vision, 2003.

[36] J. Wu and J.M. Rehg, "Where am I: Place instance and category recognition using spatial PACT," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2008.

[37] J.J. Kivinen, E.B. Sudderth, and M.I. Jordan, "Learning multiscale representation of natural scenes using dirichlet processes," Proc. IEEE Int'l Conf. Computer Vision, 2007.

[38] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2005.

[39] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," IEEE Trans. Pattern Anal. Mach. Intell., vol.27, no.10, pp.1615–1630, 2005.

[40] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," IEEE Trans. Pattern Anal. Mach. Intell., vol.20, no.3, pp.226–239, 1998.

Aram Kawewong received the B.S. degree in Computer Engineering from Chulalongkorn University in 2005, and M.S. degree from Tokyo Institute of Technology in 2008. He is currently the PhD student of Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology. His research focuses on Visual SLAM and Robotic Vision.

Sirinart Tangruamsub received the B.S. degree in Computer Engineering from Chulalongkorn University in 2005, and M. Eng. degree from Chulalongkorn University in 2008. She is currently the PhD student of Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology. Her research focuses Computer Vision and Image Processing.

Osamu Hasegawa received his Dr. Eng. degree in electronic engineering from the University of Tokyo in 1993. He was a research scientist at Electrotechnical Lab (ETL) from 1993 to 1999 and at Advanced Industrial Science and Technology (AIST) from 2000 to 2002. From 1999 to 2000 he was a visiting scientist at the Robotics Institute, Carnegie Mellon University. In 2002 he became a faculty member of the Imaging Science and Engineering Lab, Tokyo Institute of Technology. In 2002, he was jointly appointed researcher at PRESTO, Japan Science and Technology Agency (JST). He is a member of the IEEE Computer Society and IPSJ.