Robotics and Autonomous Systems 96 (2017) 224–237

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Biologically-inspired visual place recognition with adaptive multiple scales

Chen Fan [a,b,*], Zetao Chen [c], Adam Jacobson [b], Xiaoping Hu [a], Michael Milford [b]

[a] College of Mechatronic Engineering and Automation, National University of Defense Technology, China
[b] Australian Centre for Robotic Vision, Queensland University of Technology, Australia
[c] Department of Mechanical and Process Engineering, Swiss Federal Institute of Technology in Zurich, Switzerland

HIGHLIGHTS

- A clustering-based algorithm for adaptively calculating the spatial scales.
- Specific scale-encoding rules to model the grid cells' encoding patterns.
- A novel coarse-to-fine framework for recognizing places at multiple scales in a parallel manner.
- A comprehensive analysis of how and why the significant performance. improvements are achieved using this new approach.

ARTICLE INFO

Article history:
Received 18 December 2016
Received in revised form 22 June 2017
Accepted 28 July 2017
Available online 8 August 2017

Keywords:
Visual place recognition
SLAM
Adaptive multiple scales
Metric learning
Biologically-inspired
Grid cells

ABSTRACT

In this paper we present a novel *adaptive* multi-scale system for performing visual place recognition. Unlike recent previous multi-scale place recognition systems that use manually pre-fixed scales, we present a system that adaptively selects the spatial scales. This approach differs from previous multi-scale methods, where place recognition is performed through a non-optimized distance metric in a fixed and pre-determined scale space. Instead, we learn an optimized distance metric which creates a new recognition space for clustering images with similar features while separating those with different features. Consequently, the method exploits the natural spatial scales present in the operating environment. With these adaptive scales, a hierarchical recognition mechanism with multiple parallel channels is then proposed. Each channel performs place recognition from a coarse match to a fine match. We present specific techniques for training each channel to recognize places at varying spatial scales and for combining the place recognition hypotheses from these parallel channels. We also conduct a systematic series of experiments and parameter studies that determine the effect on performance of using different numbers of combined recognition channels. The results demonstrate that the adaptive multi-scale approach outperforms the previous fixed multi-scale approach and is capable of producing better than state of the art performance compared to existing robotic navigation algorithms. The system complexity is linear in the number of places in the reference static map and can realize the online place recognition in mobile robotics on typical dataset sizes We analyze the results and provide theoretical analysis of the performance improvements. Finally, we discuss interesting insights gained with respect to future work in robotics and neuroscience in this area.

1. Introduction

Visual place recognition techniques operating at one fixed spatial scale, or over two, achieve impressive performance in robotic mapping and localization tasks [1–3]. Recent high profile discoveries in neuroscience have demonstrated that animals such as rodents, and likely many other mammals including humans, can navigate the world using multiple parallel maps, each of which encodes the world at varying spatial scales [4]. Unlike hybrid metric-topological multi-scale robot mapping systems, rodent maps are homogeneous, distinguishable only by the scale [3]. These multiple scales encoded in rodent neurons can represent the spatial area ranging from a few square centimeters to several square meters, with many intermediate scales represented in-between. This multi-scale representation has been theoretically proven to be beneficial for efficient mapping of arbitrarily large environment [5,6].

* Corresponding author at: College of Mechatronic Engineering and Automation, National University of Defense Technology, China.
E-mail address: fanchen.nudt@hotmail.com (C. Fan).

http://dx.doi.org/10.1016/j.robot.2017.07.015
0921-8890/

In robotics, several recent investigations have demonstrated that a multi-scale mapping system improves the performance of robotic navigation, including using the heterogeneous and homogeneous scales. Heterogeneous multi-scale systems [1,2,7] generally use hybrid mapping frameworks with a combination of a local metric map and a global topological map to achieve real-time large-scale navigation. The heterogeneous mapping frameworks are typically limited to two distinct scales. In contrast, a recent homogeneous multi-scale system [3,8] inspired by the rodent multi-scale map utilizes an array of spatially specific place recognition networks, with each one trained to perform place recognition at a specific spatial scale. The final place match is generated by combining the place recognition hypotheses from these multiple spatial scales. The spatial scales in this approach are homogeneous and driven by the associations between the sensor input and the environment. Although there is little obvious evidence that the homogeneous scales or heterogeneous scales are better for place recognition, this study focus on investigating the potential benefits of adaptively homogeneous scales on place recognition in challenging real world environment.

In this research, we consider the homogeneous multi-scale system for robotics and aim to build an optimal association between the sensor input and the environment. Unlike previous approaches using manually pre-fixed scales, our approach adaptively calculates the spatial scales and maps the sensor input to a feature-based space in which places are recognized at varying spatial scales. We present four key novel contributions:

- a clustering-based algorithm for adaptively calculating the spatial scales,
- specific scale-encoding rules to model the grid cells' encoding patterns,
- a novel coarse-to-fine framework for recognizing places at multiple scales in a parallel manner, and
- a comprehensive analysis of how and why significant performance improvements are achieved using this new approach.

These contributions contribute to exploit the nature spatial scales in varying environments and build the parallel processes for recognizing places at adaptively multiple scales. We conduct experiments on two robotics benchmark datasets and compare our method with other state-of-the-art place recognition algorithms. The results demonstrate that using adaptive multiple scales leads to a significantly improvement in place recognition performance. Compared with the previous fixed multi-scale algorithm [3], the recall rate at 100% precision can be improved from 47% to 78% (Precision is the proportion of correctly retrieved places, and recall is the fraction of correctly retrieved places out of all the correct places in the dataset. A perfect system would be the one that achieves a precision of 100% and a recall rate of 100%).

This paper is organized as follow. Section 2 briefly discusses related place recognition and mapping techniques. In Section 3, we describe the key adaptive multi-scale place recognition algorithm. Section 4 presents the detailed experiments and results with discussions. The conclusions and future work are provided in Section 5.

2. Related work

In robotics, place recognition – the ability to recognize places that the robot has visited before – plays an important role in many robotic research fields including simultaneous localization and mapping (SLAM), localization, mapping, and recognition. Over the past four decades, there have been a number of algorithms proposed which only utilize one fixed scale, including FAB-MAP [9,10], MonoSLAM [11], RatSLAM [12–15], FrameSLAM [16] and SeqSLAM [17,18]. Multi-scale mapping approaches have also been proposed, which often take the form of a hybrid metric-topological or topometric map [1,2,7,19–21]. These typical approaches develop a metric representation of the local environment, with the large-scale space being represented by a topological map. The metric map in these approaches is often in form of the Euclidean map. Atlas [2] is a hybrid SLAM algorithm which combines existing small-scale mapping algorithms with a global topology to achieve real-time large-scale navigation. Similarly, a hybrid extension to the Spatial Semantic Hierarchy [7] builds local maps using metric SLAM methods but represents the structure of large-scale space using a topological map. Different types of maps are used at each scale, and these methods by which they are combined are typically bespoke techniques. Furthermore, the spatial scales of these hybrid systems are heterogeneous (metric and topological) and limited to two distinct scales. For generating the topological map, the LexToMap [21] associates the topological locations to the image annotations through a pre-trained Convolutional Neural Networks (CNN [22] model. Recent place recognition research inspired by rodent neural maps [23] has started to model the multiple homogeneous spatial scales encoded in the mammalian brain and demonstrated that state-of-the-art performance can be achieved [3].

Extensive research has been conducted on the mapping and navigation mechanisms in rodents, leading to the discoveries of different spatial cognitive cells [24–26]. Recently, a new type of spatial encoding cells, called a grid cell, was found within the mammalian brain, including rats, monkeys and bats [4,27–29]. The grid cell is closely related to the animals' mapping and navigation activity, whose firing structure reveals the characteristics of multiple discrete and overlapping scales. Grid cells can fire maximally whenever the animal is located at the vertices of a regular grid of equilateral triangles over the environment. Plotting the spatial autocorrelogram of the neural activity of a grid cell reveals the triangular tessellating nature of its firing field [4]. When we define the scale of a grid cell as the distance between each place fields, the whole population of grid cells encode space with multiple, discrete scales [23]. Furthermore, grid cells in the same area of the brain fire with the same spacing and orientation, but with different phasing, and together cover every point of the environment (the overlapping patterns). Fig. 1 illustrates the autocorrelogram of neural activity of three simulated grid cells, where the scales of the three place fields increase from left to right. The area encoded by a cell at each grid vertex can vary from a few square centimeters to ten of square meters, with an unknown upper limit. Studies have demonstrated in theory that the integrated, multi-scale grid maps are beneficial for efficient mapping of arbitrarily large environments [5,6].

Many biologically-inspired place recognition systems have been proposed in robotics which try to mimic the discovered mapping and navigation mechanisms in animal brains. The homogeneous multi-scale system [3] – a representative of paradigms which is most closely related to the approach in this paper – is based on the discrete multi-scale structure of grid cells in rodent hippocampus [23], and can utilize arrays of distance metrics, with each one trained to perform place recognition at a manually fixed spatial scale. These fixed scales hardly mimic the actual neural activity of grid cells in a varying environment. Furthermore, the manually pre-determined scales can damage the natural groups in training the distance metric. Although the experimental results show good place recognition performance with these fixed spatial scales, there is still potential improvement for furthering the biologically inspiration by using adaptive multiple spatial scales.

The grid cell formation within the mammalian hippocampus utilizes multiple scales, overlapping firing patterns and hexagonal representations of space which have been shown to dynamically

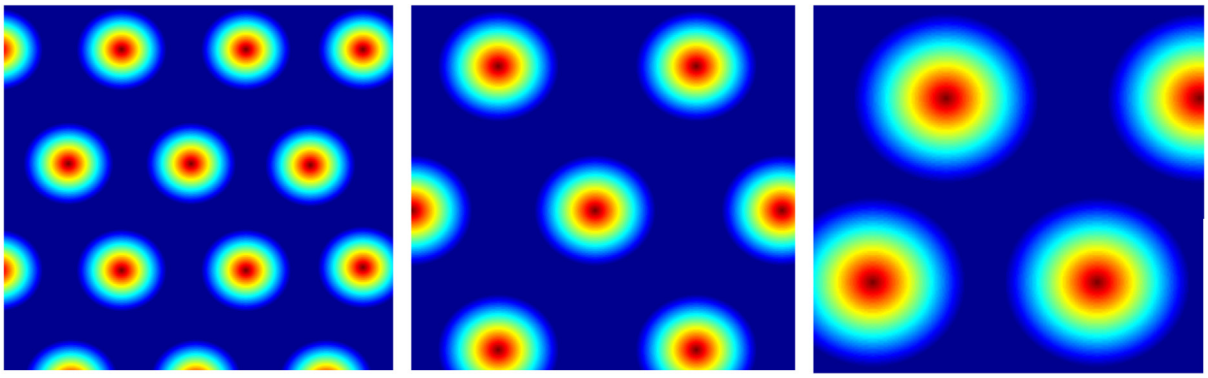

Fig. 1. Autocorrelograms of the firing fields of three simulated grid cells with varying scales. As grid firing fields become larger their encoding space also increases, maintaining the same overall field structure.

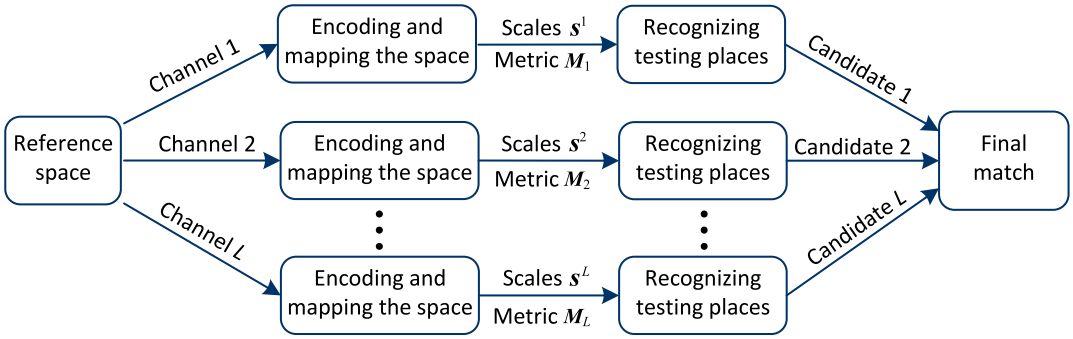

Fig. 2. Schematic of the place recognition system. Each horizontal row represents a single recognition channel in the system. These recognition channels are parallel and independent. Each channel uses a clustering-based algorithm to adaptively select the multiple spatial scales for encoding the space and a machine learning method to train an optimal distance metric for recognizing places. These resulting multiple place recognition candidates are combined to produce a final place match.

alter spatial representations based on environmental information to enable navigation within a variety of environments. This work is an initial study into specifically identifying the potential benefits of utilizing multiple place recognition channels, each of which adaptively selects spatial scales based on environmental similarity and the implementation of overlapping firing fields. Prior research [3] into multi-scale systems provided a proof of concept that place recognition performance can be significantly improved by using multiple homogeneous scales. However, the previous approach lacks the ability to adaptively calculate the scales, instead relying on manually determined fixed scales. In animals, it is still unknown whether the absolute scales are fixed or determined adaptively—this paper focuses on the development of an adaptive system that for the first time chooses spatial scales that maximize place recognition performance in a specific environment.

3. Approach

The place recognition approach presented here is inspired by the multi-scale structure of grid cells discovered within rodent brains [23]. In this section, we describe the image features used, the learning methods for adaptively selecting the spatial scales and recognizing places at varying spatial scales, and the mechanism of combining recognition candidates from different process channels to produce an overall matching result. A schematic of the place recognition process is shown in Fig. 2.

3.1. Extracting features

There is no specific reliance on the feature type utilized in our approach. In this paper, we evaluate two commonly used features: Grayscale intensity 'feature' and GIST features. For Grayscale features, the image is firstly down-sampled to 48 × 64 pixels, and then Principal Component Analysis (PCA) [30] is applied. We select the top 500 principal dimensions as the image representation which captures approximately 90% of the data variance. For GIST features, we use the shape model proposed by Oliva and Torralba [31], GIST features are extracted from these down-sampled images in a 512-dimensional vector. Then PCA is applied once again, selecting the top 300 principal eigenvalues as the image GIST representation, which encodes 90% of the data variance. The captured percentage has been found to be sufficient, with the image feature reduction from PCA having minimal negative effect on the final results.

3.2. Modeling grid cells

In order to model the discrete multi-scale encoding patterns of grid cells [23], our place recognition system segments the environment into multiple sizes of image clusters representing multiple spatial scales. Fig. 3 demonstrates how multiple spatial scales can be extracted from a real dataset. The blue arrow indicates the route that the system is trained to recognize, and the red ellipses represent the active clusters, or grid cells, of a particular layer of our 'grid cell' model. We refer to the spatial scale by the size of the active cluster, i.e. the number of images in the red ellipse. The images that are captured with a particular active cluster have similar scene or environmental characteristics and are all assigned the same label (the number on the red ellipses). The sizes of these active clusters are different, which correspond the multiple encoding spatial scales of grid cells. All frames that are captured outside any red ellipses are considered negative as samples and are assigned to another cluster label. These labels are used for training an optimized distance metric in place recognition.

Fig. 3. Encoding patterns example for grid cells in a real dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

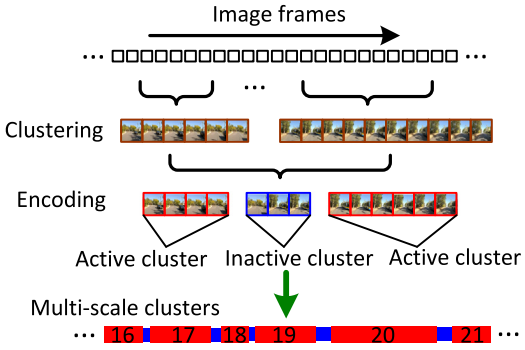

Fig. 4. There are two steps involved in encoding image frames at multiple scales. Firstly, we use cluster analysis to group the image frames into a number of clusters with different lengths. Then, according to the encoding rules, these clusters are divided into active (red area) and inactive (blue area). The active cluster is captured in a region that has similar scene or environmental characteristics (red ellipses in Fig. 3), which corresponds the firing neural activity of that grid cells. The inactive clusters include all the negative samples from the route. The size of an active cluster represents a specific spatial scale of that grid cell. The numbers on the red rectangles denote the label of active clusters for training a distance metric. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Encoding space

The encoding patterns in the rodent brain using multiple discrete spatial scales have shown a number of theoretical advantages for mapping of large environments [5,6]. Inspired by the studies, the encoding space is proposed to adaptively obtain spatial scales. Firstly we use the cluster analysis to adaptively obtain natural image clusters, and then propose encoding rules to produce multiple discrete spatial scales. Fig. 4 shows the process of encoding in a single recognition channel.

3.3.1. Image clustering

In this paper, we use the clustering algorithm to automatically identify the natural grouping of space to produce consecutive clusters with different lengths, analogous to the multiple spatial scales of grid cells. An image cluster with a longer sequence corresponds a grid cell with a larger scale. The system uses the Fuzzy c-means (FCM) method [32–34] for clustering, which has low computational complexity and is relatively effective for processing large amounts of data [35].

The key innovation of FCM clustering is to group each image into two or more clusters and identify the optimal cluster in which the images are most similar. Each cluster can be denoted as:

$$\boldsymbol{x}_i \in \boldsymbol{R}^N \sim \{\boldsymbol{c}_k\}_{k=1}^{C} \tag{1}$$

where $\boldsymbol{x}_i$ denotes the feature vector in N dimensional space and $\boldsymbol{c}_k$ is the center vector of a cluster. C is the number of clusters, which can be pre-estimated by an efficient method [33] in the initializing process.

In the FCM clustering procedure, the classification of an image is described with a dissimilarity function as below:

$$J_i = \sum_{k=1}^{C} \mu_{ik}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2 \tag{2}$$

where μ_{ik} is the membership grade of $\boldsymbol{x}_i$ in the kth cluster and m is a positive exponent for controlling the degree of fuzzy overlap. The membership grade specifies the degree of an image belonging to a cluster. An image with a higher membership grade will be more likely to reside in the cluster. The purpose of FCM clustering is to minimize the dissimilar function for obtaining the membership grade with the highest value.

Supposing there are K images in total, the objective dissimilar function is calculated with Eq. (2):

$$\text{Minimize} \quad J = \sum_{i=1}^{K} J_i. \tag{3}$$

The objective function is solved through an iterative optimization in [32]. Then the partitioning matrix $\boldsymbol{\mu}$ (the size is $K \times C$) is carried out for specifying the clusters of all images.

Fig. 5(a) gives the clustering results on the St. Lucia dataset (this dataset will be described in Section 4.1). As can be seen, all image frames can be automatically grouped into a number of clusters with different sizes, each of which stores similar images from a spatially approximate region.

3.3.2. Encoding image clusters

This section presents the encoding rules to produce the multiple spatial scales. In each recognition channel, the image clustering results are encoded into active and inactive patterns. The active cluster represents the place where the particular cluster or grid cell fires, with a particular size representing the spatial scale of that grid cell. The inactive cluster is the set of all negative samples (blue area in Figs. 3 and 4). The encoding rules used in our paper are described as follows, and the examples are illustrated in Fig. 5(b).

Rule 1: An image cluster whose size is less than the spatial scale lower limit is assigned as an inactive cluster. The spatial scale lower limit represents the minimum precision of the place recognition candidate that the system can produce, which is defined as below.

Rule 2: All the remaining image clusters are considered as the active clusters and tagged with different labels.

The spatial scale lower limit determines the accuracy that the recognition system can achieve. A cluster with smaller size will have higher recognition accuracy. For rule 1, we set the smallest spatial scale size to 4 frames in the St. Lucia dataset and 6 frames in the Eynsham dataset, which correspond to about 2 and 40 m in physical space (consistent with localization tolerances reported in other research on these datasets [3,10,36]), respectively. Rule 2 specifies that we use different labels to indicate different active clusters and assign the same label to all inactive area.

Through the process of image clustering and encoding, each recognition channel produces many consecutive active clusters with different sizes (Fig. 5(b)). The size of each active cluster corresponds the spatial scale of that grid cells. These active clusters with the same size represent that grid cells encode space at one spatial scale. Different sizes of active clusters correspond to multiple spatial scales of that grid cells encoding space. In each

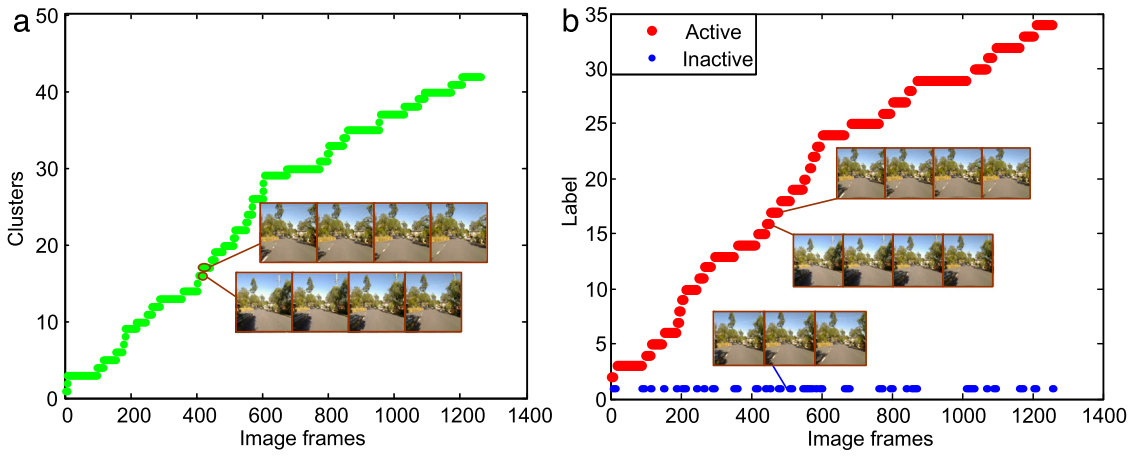

Fig. 5. Illustration of scale encoding on the St. Lucia dataset using GIST features: (a) Grouping images into different clusters. (b) Encoding clusters with different labels for training the distance metric. In figure (a), green circles represent the clustering results, each of which captures similar images from a spatially approximate region. In figure (b), the red and blue circles are the results of encoding. Red circles represent the active cluster of that grid cells, and blue circles are considered as the inactive area. The size of each active cluster represents a specific scale that grid cells encode the spatial space. The image frames that are captured within an active cluster are all assigned to the same label, and the frames in any inactive area are assigned to another label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

recognition channel, these spatial scales are described as a scale vector $\boldsymbol{s}^p$ (Eq. (4)). The labels on active and inactive clusters are used for training a distance metric ($\boldsymbol{M}^p$) in the machine learning method described below.

$$\boldsymbol{s}^p = \left[s_1^p, s_2^p \dots, s_{n_p}^p\right] \tag{4}$$

where p denotes the pth recognition channel and n_p is the number of the spatial scales.

3.4. Modeling a grid map

In this section, we use a supervised learning method, Large Margin Nearest Neighbors (LMNN) [37], to learn an optimized distance metric in each recognition channel. Each learning metric maps the data to a new space where the grid map is modeled and places can be recognized at multiple spatial scales. In each recognition channel, labels on each frame (red and blue circles in Fig. 5(b)) are used by the LMNN process involving multiple scales which maps new incoming data to a space such that the k nearest neighbors of any frame inside a cluster always come from the same cluster.

The metric training procedure consists of two steps. The first step involves identifying a number of k nearest neighbors for each input $\boldsymbol{x}_i$. We specify the target neighbors of an image by its nearest neighbors in the same cluster. The second step is to train a Mahalanobis distance metric $\boldsymbol{M}^p$ (the pth channel) such that all target neighbors of a frame $\boldsymbol{x}_j$ are closer to $\boldsymbol{x}_i$ than any other samples with different labels. A Mahalanobis distance metric $\boldsymbol{M}^p$ computes the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as:

$$d_p\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^T \boldsymbol{M}^p \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right) \tag{5}$$

where $\boldsymbol{M}^p \geq 0$ should be a semidefinite matrix to generate a positive distance measurement. It can be calculated by:

$$\boldsymbol{M}^p = \boldsymbol{v}^T \boldsymbol{v} \tag{6}$$

where $\boldsymbol{v}$ is an arbitrary real $N \times N$ matrix.

Substituting Eq. (6) into Eq. (5) results in:

$$d_p\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^T \boldsymbol{v}^T \boldsymbol{v} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right) = \left\| \boldsymbol{v}\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)\right\|^2 \tag{7}$$

where the transformation vector $\boldsymbol{v}$ maps the data to a new space in which Euclidean distance is calculated.

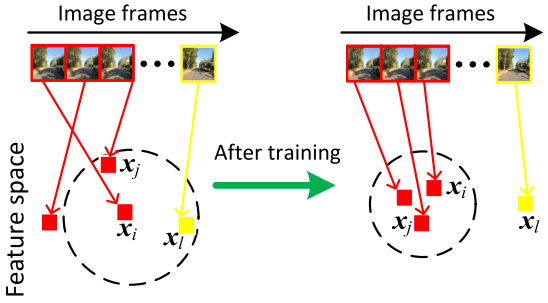

Fig. 6. Images distributed in the feature space before (left) and after (right) training. The red boxes represent the images , which come from the same cluster and share the same labels. The yellow box represents the images from a different cluster. Before training, the yellow feature $\boldsymbol{x}_l$ invades the boundary (circle dotted line) of the red feature $\boldsymbol{x}_i$ defined by its neighbor $\boldsymbol{x}_j$ (using the Euclidean distance). Ideally after training, all image features with the same label lie within a smaller radius and the differently labeled feature is separated. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We use $y_{ij} \in \{1, 0\}$to indicate whether or not $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ share the same label. The $\varepsilon_{ijl} \geq 0$ indicate the amount by which a differently labeled sample $\boldsymbol{x}_l$ invades the boundary of $\boldsymbol{x}_i$ defined by all its neighbors $\Sigma_{j \to i}\boldsymbol{x}_j$ where the notation $j \to i$ indicates instance $\boldsymbol{x}_j$ is a target neighbor of instance $\boldsymbol{x}_i$. The training process is illustrated in Fig. 6. Using the semidefinite program, the metric $\boldsymbol{M}^p$ is given by:

$$\textit{Minimize} \quad \sum_{j \to i}\left[d_p\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + \lambda \sum_l \left(1 - y_{il}\right)\varepsilon_{ijl}\right] \tag{8}$$

subject to

(1) $d_p\left(\boldsymbol{x}_i, \boldsymbol{x}_l\right) - d_p\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \geq 1 - \varepsilon_{ijl}$

(2) $\varepsilon_{ijl} \geq 0$

(3) $\boldsymbol{M}^p \geq 0$

where the weighting parameter $\lambda \in [0, 1]$ balances the two terms in the objective function and can be tuned via cross-validation.

The constraint (1) penalizes the differently labeled input $\boldsymbol{x}_l$, which invades the local neighbors of $\boldsymbol{x}_i$. The constraint (2) enlarges the feasible region with the slack variable ε_{ijl} and the constraint (3) enforce the $\boldsymbol{M}^p$ to be positive definite. Since the distance $d_p\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$

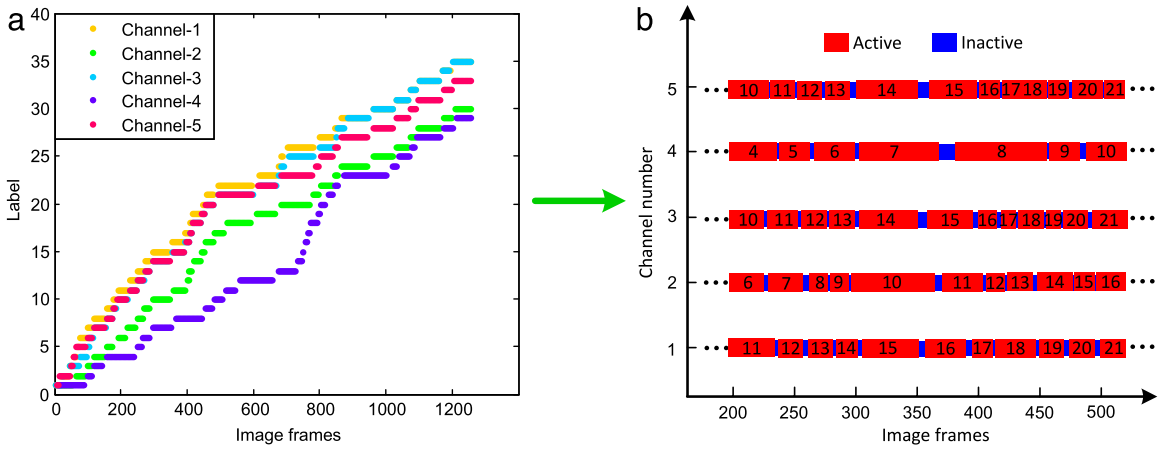

Fig. 7. Multiple recognition channels in a real dataset. Figure (a) shows the encoding results in five channels. In each channel, the encoding result consists of many short dashed lines, which represent different sizes of image clusters, corresponding to the spatial scales of grid cells encoding space. The number and size of the clusters in these channels differ from each other. Each image is assigned to five different active clusters to create overlapping encoding patterns, as occurs with grid cells. It means that one place can be encoded at different scales in our system. Figure (b) illustrates the multi-scale and overlapping encoding results in these recognition channels. The red rectangles represent the active clusters of that grid cells. Blue sections are the area where the grid cells are inactive. The size of each active cluster represents a specific scale that grid cells encode the space. The numbers on the active clusters denote the labels (corresponding to the number on the y-axis in figure (a)) of clusters which are used for training the distance metric. The multi-size active clusters in each channel represents the multi-scale structure of grid cells. Across these channels, each image is encoded at different scales to create the overlapping patterns. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is linear in the matrix $\boldsymbol{M}^p$, Eq. (8) is a semidefinite program and a global minimum can be efficiently computed.

3.5. Multiple recognition channels

There are multiple place recognition channels (in total L) in the proposed system, each of which encodes places at multiple spatial scales. The encoding patterns are overlapping across these channels, which corresponds that the grid cells from different neural networks fire in an overlapping way. Fig. 7 demonstrates the multi-scale and overlapping patterns with five recognition channels in a real dataset. Instead of using different time offsets to mimic this overlapping patterns [3], we vary the degree of fuzzy overlap between clusters in FCM [34] to generate different clustering results in each recognition channel. The multi-size image clusters in each channel denote the multiple encoding scales of grid cells. Across these channels, each place can be encoded at different spatial scales to create overlapping patterns. In each channel, the different spatial scales and the distance metric are learned independently. Assuming there are L place recognition channels and for each channel the spatial scales are denoted as a scale vector $\boldsymbol{s}^p$ and the distance metric is $\boldsymbol{M}^p$. These different scale vectors and the distance metrics from each recognition channels are provided in the form of sets:

$$\boldsymbol{s} = \left\{\boldsymbol{s}^1, \boldsymbol{s}^2 \cdots, \boldsymbol{s}^L\right\} \tag{9}$$

$$\boldsymbol{M} = \left\{\boldsymbol{M}^1, \boldsymbol{M}^2 \cdots, \boldsymbol{M}^L\right\}. \tag{10}$$

3.6. Place recognition

In this section, we describe a hierarchical coarse-to-fine strategy for recognizing places in each place recognition channel. The spatial scales in the coarse-to-fine strategy are derived from the image encoding results. Using the pth channel as an illustration, the multiple spatial scales are in form of a scale vector $\boldsymbol{s}^p$ and the distance metric is $\boldsymbol{M}^p$ (Fig. 2). The smallest scale $s^p_{\min}$ is used for fine matching and each of the other $\left(n_p - 1\right)$ larger scales is used for coarse matching. Each of the $\left(n_p - 1\right)$ larger scales generates a coarse place recognition estimate within which a finer match

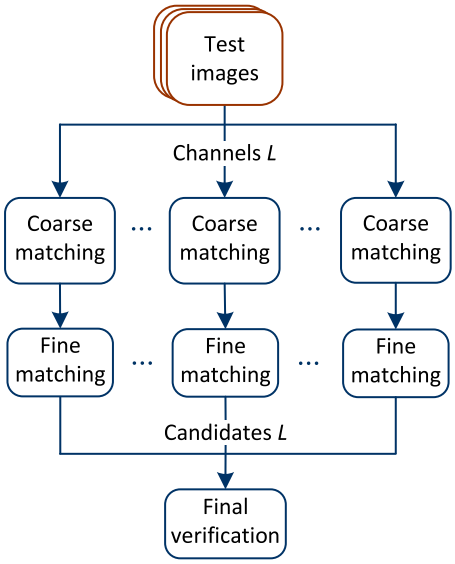

Fig. 8. Schematic of place recognition across multiple channels. Each channel performs a coarse-to-fine matching and these resulting place recognition candidates are combined to produces a final place match.

at scale $s^p_{\min}$ is used to produce a hypothesis. Combining these $\left(n_p - 1\right)$ hypotheses, we choose the one with lowest score as the matching candidate of each channel. Since there are L different place recognition channels, L matching candidates will be produced. Then we combined matching candidates from these channels and pick the place match with the lowest firing score as the final matching result. The above recognition process is illustrated in Fig. 8.

3.6.1. Multi-scale matching in each channel

The multi-scale matching in each channel is performed by coarse matching at each of the larger spatial scales followed by fine matching at the smallest spatial scale (Fig. 9). Here we use the pth channel as an example to illustrate the multi-scale matching process, other channels perform a similar process. In the pth channel, we use the $s^p_{\min}$ to represent the smallest spatial scale in

the scale vector $\boldsymbol{s}^p$, each of the larger scales being denoted as s_c^p. For the coarse matching stage, the testing image is first grouped into a temporally consecutive cluster with a larger scale s_c^p and images in that cluster are denoted as $P_i \ \left(i = 1, 2 \ldots, s_c^p\right)$. All the training images are denoted as $T_j \ (j = 1, 2 \cdots T)$ and the image difference between P_i and T_j calculated using the distance metric $\boldsymbol{M}^p$ is denoted as $D_M^p \ (i, j)$. For each testing cluster, coarse localized matching is performed through the whole training space (Fig. 9) to search for the coarse place field H that best matches the testing cluster. The difference score is given by:

$$F\ (q) = \sum_{j=q}^{j=q+s_c^p-1} D_M^p\ (j-q+1, j)\,, \quad \forall q \in \left[1, T - s_c^p + 1\right] \tag{11}$$

where q is the searching variable in training space. The coarse matching field H can be obtained by the minimization function:

$$H = \arg\min_q F\ (q)\,. \tag{12}$$

The corresponding firing score is denoted as $F\ (H)$. A smaller value of $F\ (H)$ indicates a stronger and more confident match. The followed fine localized matching is performed at the smallest scale $s_{\min}^p$ in the training space $\left[H, H + s_c^p - 1\right]$. The difference score is carried out by:

$$F\ (t) = \sum_{j=t}^{j=t+s_{\min}^p-1} D_M^p\ (j-t+1, j)\,, \quad \forall t \in \left[H, H + s_c^p - s_{\min}^p\right]. \tag{13}$$

Then a fine matching hypothesis is provided by:

$$Y = \arg\min_t F\ (t)\,. \tag{14}$$

The corresponding firing score of the fine matching is $F\ (Y)$. Since there are $\left(n_p - 1\right)$ larger scales for coarse matching, $\left(n_p - 1\right)$ matching hypotheses will be produced, with the fine matching candidate chosen as the one with the lowest firing score:

$$Y_o^p = \arg\min_m \left(F\ (Y_m)\right), \quad \forall m \in \left[1, n_p - 1\right] \tag{15}$$

with the fine localization (the matching candidate of the pth recognition channel) at the smallest scale in training space: $\left[Y_o^p, Y_o^p + s_{\min}^p - 1\right]$.

3.6.2. Final verification

There are L different process channels for recognizing places in parallel way. These channels generate L different matching candidates. To determine the most likely place match, we combine these candidates and pick the one with the lowest matching scores:

$$Y_o\ (final) = \arg\min_p \left(F\left(Y_o^p\right)\right), \quad \forall p \in [1, L]\,. \tag{16}$$

4. Experiments and results

In this section, we describe the experiments and results, including the datasets used and the clustering and encoding results of each recognition channel. The place recognition performance is compared with other state-of-the-art place recognition algorithms, including multi-scale (MS) place recognition [3], SeqSLAM [18] and FABMAP [9,10]. We also present an analysis evaluating the performance improvement from using the adaptive multi-scale (AMS) place recognition system and a parameter study to determine the effect on performance of using different numbers of combined recognition channels.

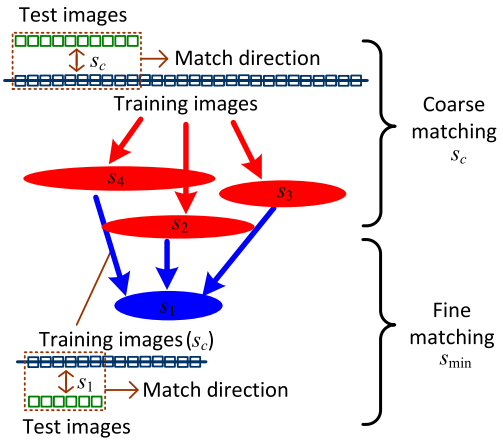

Fig. 9. The coarse-to-fine matching process in a place recognition channel. Each of the three larger scales (s_2, s_3, s_4) provides a coarse place recognition estimate (red ellipse), followed by the fine matching (blue arrow) at the smallest scale s_1. The hypothesis with the smallest scale is picked as the matching candidate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1. Datasets

We used two benchmark datasets in our experiments with the dataset routes showed in Fig. 10. Both datasets consist of two traverses along the same routes. Videos from the first traverse are used for training while videos from the second traverse are used for testing. The training database is static and built off line. The place recognition is an online process using the static map. The first dataset comprises of a forward facing monocular camera videos from a car traveling along a selection of streets in the suburb of St. Lucia, Brisbane (Fig. 10(a)). The videos are captured at 640 × 480 pixel resolution at an average frame rate of 15 frames per second. The datasets consist of about 2000 images and two traverses—the first traverse is gathered in the morning and the other in the afternoon. The odometry information is simulated by linearly interpolating the GPS data (1 Hz) to drive the cluster formulation at about 0.67 m/frame. The second dataset is the Eynsham dataset (Fig. 10(b)) which is a large 70 km road-based dataset (2 × 35 km traverses). The images come from an omnidirectional camera (Ladybug 2) and are captured at 7 m intervals at size of 256 × 240 pixel resolution. The dataset provides GPS-derived ground truth. The two datasets are also used in [3], enabling comparison to the previous fixed multi-scale approach.

4.2. Experimental procedure

Experiments consisted of five steps: extracting features, clustering images, encoding clusters, training metrics and recognizing places. Firstly, Grayscale and GIST features were extracted from all images and used for evaluating performance respectively. Secondly, the images from the first traverse were grouped into different clusters. Thirdly, these image clusters from each recognition channel were encoded according proposed rules. In our paper, we refer to the spatial scale by the number of frames in each active cluster. For example, a spatial scale of 10 means that the active cluster in that scale contains 10 frames. The smallest spatial scale represents the upper precision of place match that the system can produce, which is pre-defined as the same for all place recognition channels. For the St. Lucia dataset, we choose the smallest spatial scale to be 2 m which corresponds 4 frames. For the Eynsham dataset, the smallest scale is about 40 m which corresponds 6 frames. The fourth step involved training a distance

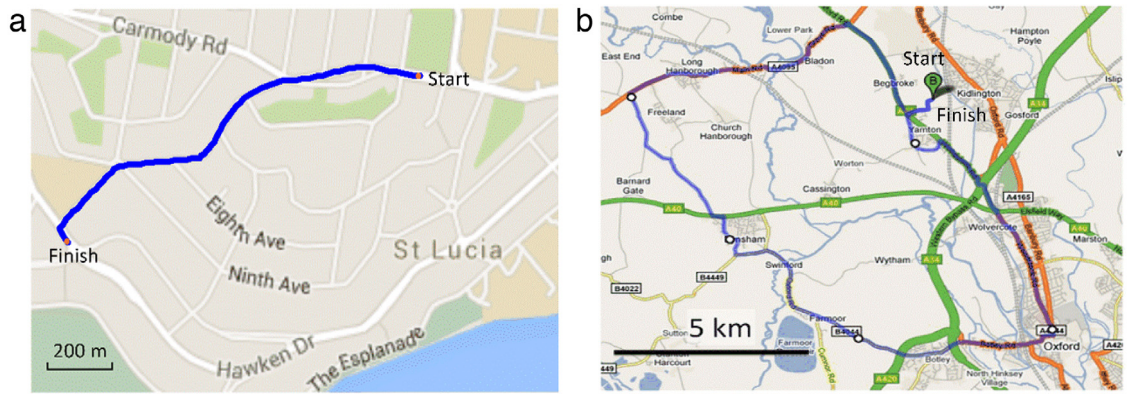

Fig. 10. Aerial images showing the dataset routes from the (a) St. Lucia and (b) Eynsham datasets
Source: Google maps.

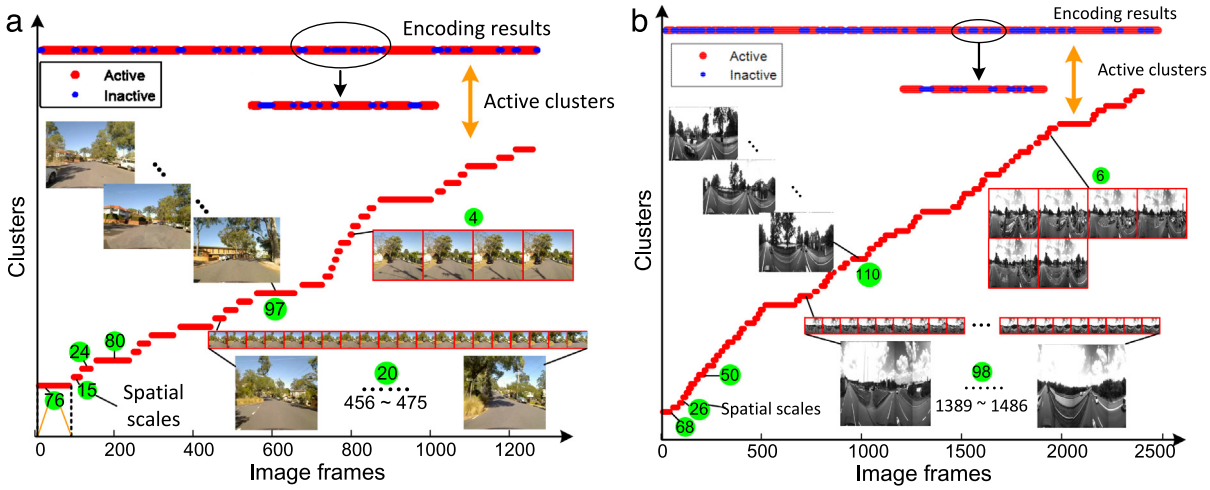

Fig. 11. Illustration of clustering and encoding results using GIST features on the St. Lucia dataset (a) and Eynsham dataset (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

metric at multiple spatial scales for each recognition channel. During the training, 20% of the training data was used for validation to prevent overfitting and facilitate early stopping during training. Finally, place recognition was performed with images from the second testing route traverse, accompanied by extensive performance evaluation, improvement analysis and parameters studies. The thresholds of correct matches in the St. Lucia dataset and the Eynsham dataset are 2 m and 40 m, which are consistent with the tolerance used in the previous study [3,10,36].

4.3. Results

4.3.1. Clustering and encoding results

In this section, we present the clustering and encoding results of a recognition channel using GIST and Grayscale features. Fig. 11 shows that the encoding results (red-and-blue lines at the top of figures) consist of alternating active (red) and inactive (blue) clusters with multiple sizes. Local regions in the black ellipse are zoomed to show the discrete encoding patterns. The number on the green circle is the size of the image cluster, which corresponds the scale of grid cells encoding the space. The sampled images illustrate some similar examples in these active clusters. In each recognition channel, the active clusters with multiple spatial scales are assigned by different labels for training a distance metric. It can be seen that each place recognition channel performs the encoding patterns with multiple scales, as occurred with grid cells.

4.3.2. Performance comparison

In this section, we use a common evaluation metric used for place recognition – the precision-recall (PR) curve – to evaluate the performance. Precision is defined as the number of correctly retrieved places divided by the total number of retrieved places. Recall is defined as the fraction of correctly retrieved places out of all the recallable places in the dataset. Precision-recall curve is a two-dimensional plot with the x-axis indicating recall and the y-axis indicating precision. Perfect performance occurs when the precision remains equal to one, as the recall increases from 0 to 1. In that perfect performance scenario, the area under the curve (AUC) equals 1. We vary a threshold which is compared to final matching scores to generate varying recalls with different precisions.

We present PR curves and the AUC results for FABMAP [10], SeqSLAM [18], MS [3] place recognition and AMS place recognition experiments using GIST and Grayscale features on the St. Lucia dataset (Fig. 12) and Eynsham (Fig. 13) dataset.

Compared with the previous MS system, the AMS matching significantly improves the performance. On the St. Lucia dataset (Fig. 12), the area under the curve (AUC) on the precision curve has been improved by about 26% using GIST features and 21% using Grayscale features, in absolute terms. Moreover, using Grayscale features can improve the recall rate at 100% precision from 0% to 4.5%. On the Eynsham dataset (Fig. 13), the improvement of AUC is from 90% to 95% on Grayscale features while using GIST features both MS and AMS can be very close to 1. It is notable that the recall rate at 100% precision increases from 47% to 78% by using GIST

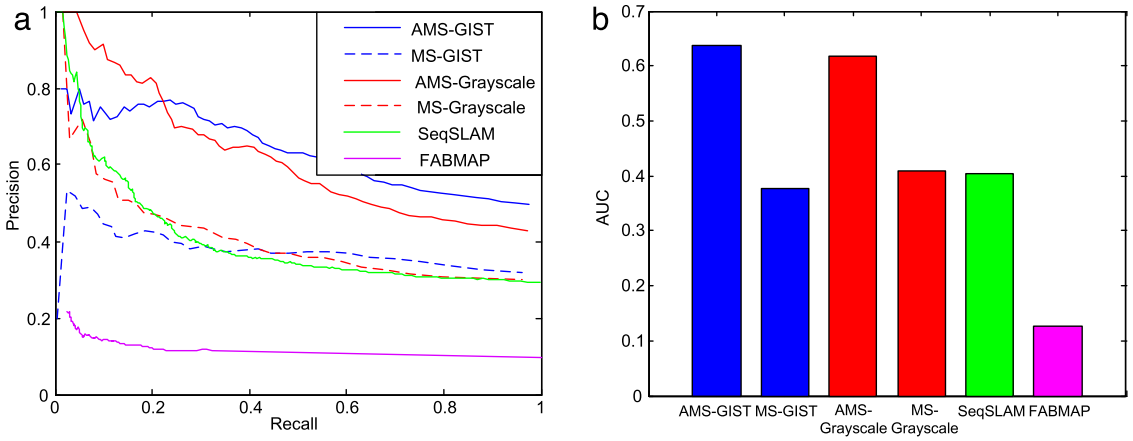

Fig. 12. Comparisons on the St. Lucia dataset using GIST and Grayscale features. (a) Precision-recall curves. (b) Area Under the curve (AUC) on the precision-recall curve.

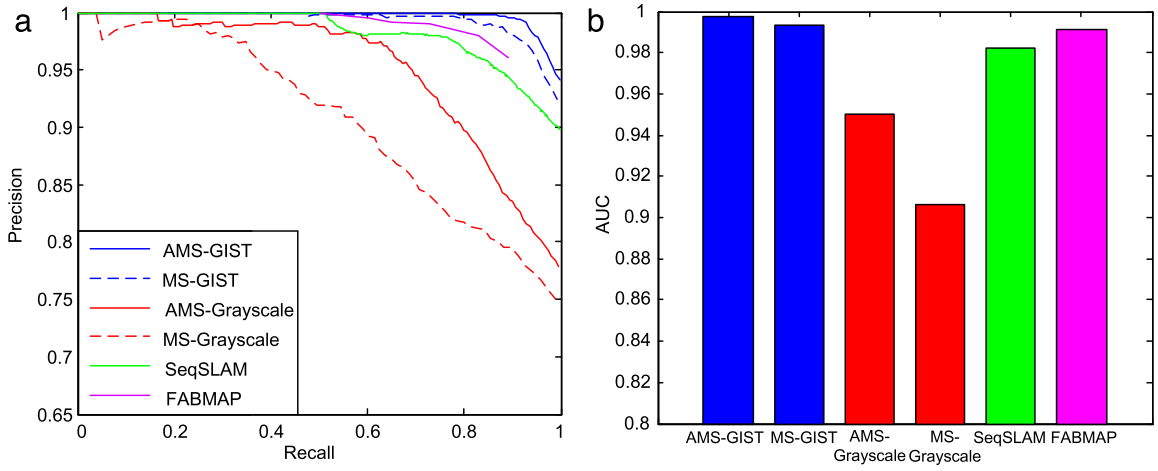

Fig. 13. Comparisons on the Eynsham dataset using GIST and Grayscale features. (a) Precision-recall curves. (b) Area Under the curve (AUC) on the precision-recall curve.

features. The corresponding improvement using Grayscale feature is from about 4% to 16%.

We also provide absolute performance comparison to other state-of-the-art systems. On the St. Lucia dataset, when using Grayscale and GIST features, the AMS system consistently outperforms FABMAP [10] and SeqSLAM [18] algorithm. The best performance on the Eynsham dataset is observed on GIST features. The maximum recall rate of about 78% at 100% precision is superior to the 49% recall rate achieved by FABMAP and 51% recall rate achieved by SeqSLAM. It also can be seen that the improvements of AMS algorithm are clearer on the St. Lucia dataset than on the Eynsham dataset. Based on the characteristics of the algorithms, it appears that the AMS method is most beneficial under conditions which undergo larger degrees of appearance-change.

4.3.3. Improvement analysis

In this section, we compare AMS and MS [3] algorithms using a single spatial scale to analyze the improvement resulting from the AMS approach. Firstly, we provide a theoretical analysis for the improvement of AMS algorithm. Both AMS and MS algorithms map the data to a new space using a distance metric, which drastically affects the place recognition performance in the new space.

The main reason for the improvement is that the AMS algorithm provides an optimized distance metric for clustering images with similar features while separating those with different features. As shown in Fig. 14, the AMS algorithm learns an optimized distance metric through these adaptive-scale clusters. These clusters

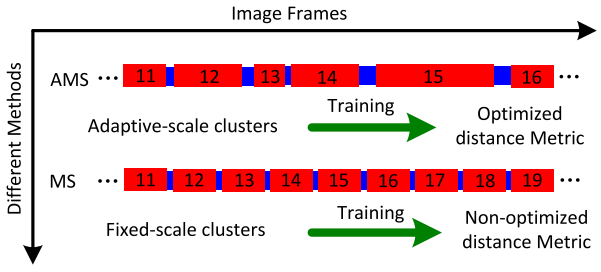

Fig. 14. Different clusters for training metric in AMS and MS.

consists of different numbers of similar images. This encoding patterns with varying spatial scales can preserve natural scene groups in varying environments. In contrast, the MS algorithm uses the manually fixed spatial scales for learning a non-optimized distance metric. These fixed scales can hardly mimic the natural activity of grid cells in varying environment and even damage the natural groups in training the distance. Fig. 15 presents the difference of image groups between the fixed multi-scale (MS) system and the adaptive multi-scale (AMS) system. Compared with the AMS system, the groups' damage of the fixed MS system is mainly reflected in two aspects: (1) all the image groups share a fixed size and this case is contradictory with the varying sizes of image groups in real varying environments; (2) the images would

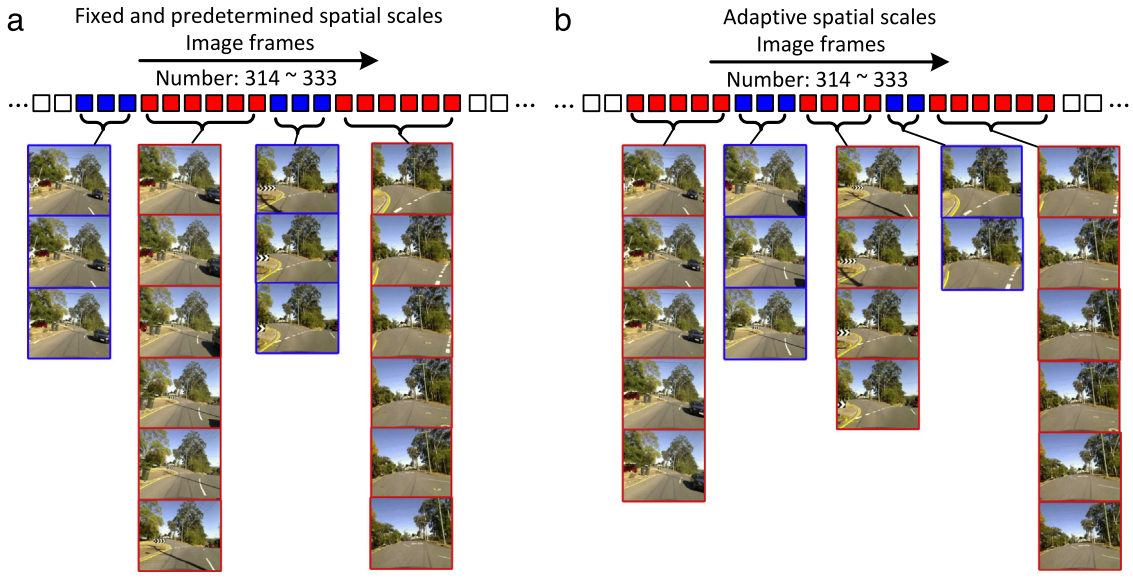

Fig. 15. Actual image groups for MS (a) and AMS (b) in the St. Lucia dataset. The red squares consist of the active image cluster and the blue squares are the inactive image cluster. The figure (a) presents the image groups with a fixed size—6 frames. The 314–320 image frame have a similar scene that a black car is always in the images. However, these images are divided into different image groups. The similar results are shown in the following group. In the figure (b), these red squares present the natural image groups in which the similar images are clustered into a group and different images are separated into different groups. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

be divided into incorrect groups, such as similar images being divided into different groups or different images being allocated to the same group. In contrast, the AMS system can adaptively cluster images with similar features while separate those with different features.

In order to evaluate the effect of distance metric, we compare the recognition performance of AMS and MS algorithms using the single scale and multiple scales. The results are shown in Figs. 16 and 17. Obviously, both on the single-scale matching and multi-scale matching, the AMS with the optimized metric (red lines and red dotted lines) outperforms the MS with the non-optimized metric (blue lines and blue dotted lines).

4.3.4. Parameter studies

The final place match is combined from different place recognition channels. Each channel involves one kind of image clustering results. In order to investigate the effect of using different image clustering results, we conduct of a series of parameter studies that using different numbers of combined place recognition channels. We conduct a total of 10 recognition channels and combine any numbers of these channels to evaluate the place recognition performance. The performance on the Eynsham dataset is evaluated using both the AUC metric and maximal recall at 100% precision, while the St. Lucia dataset is evaluated using the AUC metric only as precision does not hit 100% performance using GIST features (Fig. 16(a)). The maximal recall at 100% precision is an important criterion in place recognition because any false positive can cause a catastrophic error in the map generated by a SLAM system.

Fig. 18 evaluates the influence of the number of combined recognition channels on the performance of the St. Lucia dataset, evaluated by the average AUC. Using GIST and Grayscale features, more combined channels tends to deliver better performance, with an upper limit of 8 channels. Similar results are provided on the Eynsham dataset, as shown in Fig. 19. Since the performance of using GIST features is close to 1, the improvement of AUC using GIST features is not obvious. Fig. 20 presents the maximum recall at 100% precision evaluations on the Eynsham dataset. On both features, while more combined channels always delivers better performance, the system using GIST features outperforms using Grayscale features. The possible reason is that GIST features can capture more discriminative semantics and therefore have a better performance. Almost all of the results suggest that increasing channels is beneficial up to a certain point for achieving better place recognition performance.

4.3.5. Computational performance

In our system, the mapping process is offline and the place matching is online. For the offline mapping, the feature extraction, the image clustering and the encoding process in the reference dataset are not included in the computation analysis. For the online place matching, the primary computational overhead is the coarse matching process. The complexity of the proposed system is linear in the number of places (images) in the reference static map (reference dataset). Taking the St. Lucia dataset for illustration, when we compare a query image to the reference dataset of 1261 images, and using Grayscale features (500 dimensions), 4 recognition channels and a coarse scale of 52, nearly 131 million ($1261 \times 500 \times 4 \times 52$) comparisons are performed. A single core CPU with the rate of 1 billion single byte comparisons per second can process up to approximately 7 frames per second. The requisite computational memory is about 125 MB ($1261 \times 500 \times 4 \times 52/1024/1024$). These computational results are able to realize the online place matching on typical dataset sizes. For other systems, the MS and AMS have the same complexity and the computational results are same. The single-scale system, FABMAP and SeqSLAM, the equivalent time is about 396 frames per second ($1 \text{ billion}/(1261 \times 500 \times 4)$) and the requisite memory is 2.4 MB ($1261 \times 500 \times 4/1024/1024$). A direct technique of improving computation is to optimize our algorithm code to be parallelized and perform it on a GPU platform, which can do approximately 80 billion comparisons per second and process nearly 1000 frames per second.

Another avenue for future computational optimization is the new neuromorphic computing architectures becoming available such as SpiNNaker [38] and IBM's TrueNorth chip [39], both of which may potentially be very effective at implementing neural models of the grid cells modeled using an algorithmic implementation such as that described in this paper.

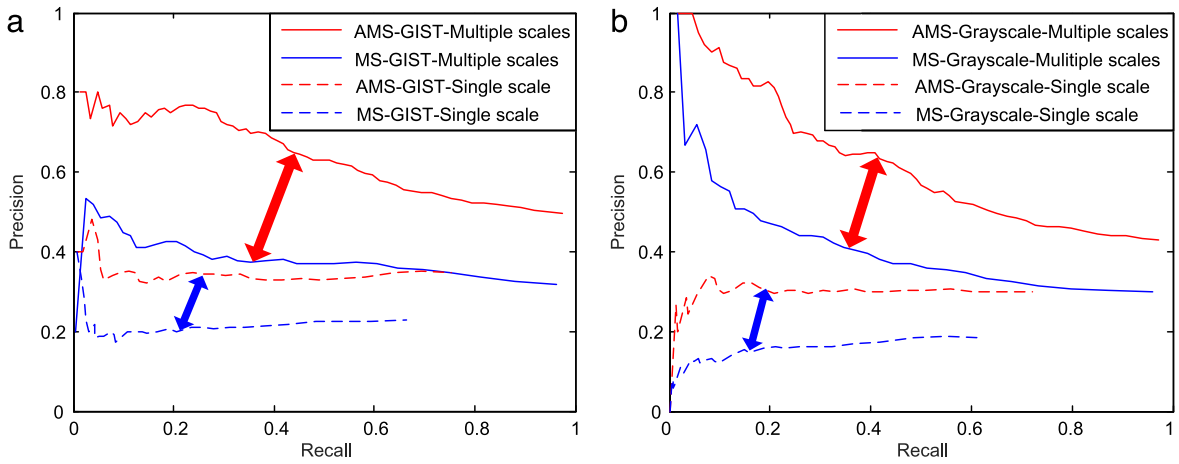

Fig. 16. Precision recall curves demonstrating the improvement between AMS and MS place recognition using GIST features (a) and Grayscale features (b) on the St. Lucia dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

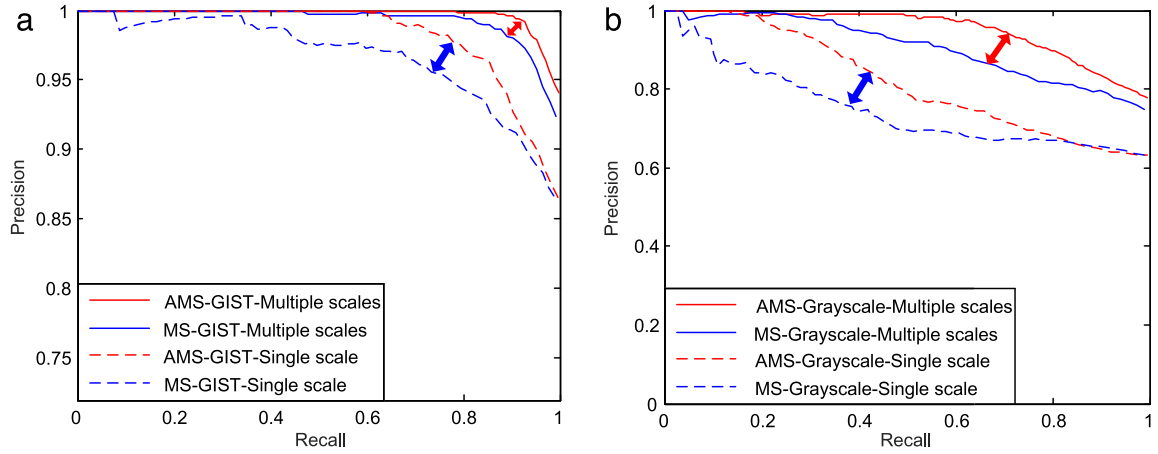

Fig. 17. Precision recall curves demonstrating the improvement between AMS and MS place recognition using GIST features (a) and Grayscale features (b) on the Eynsham dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

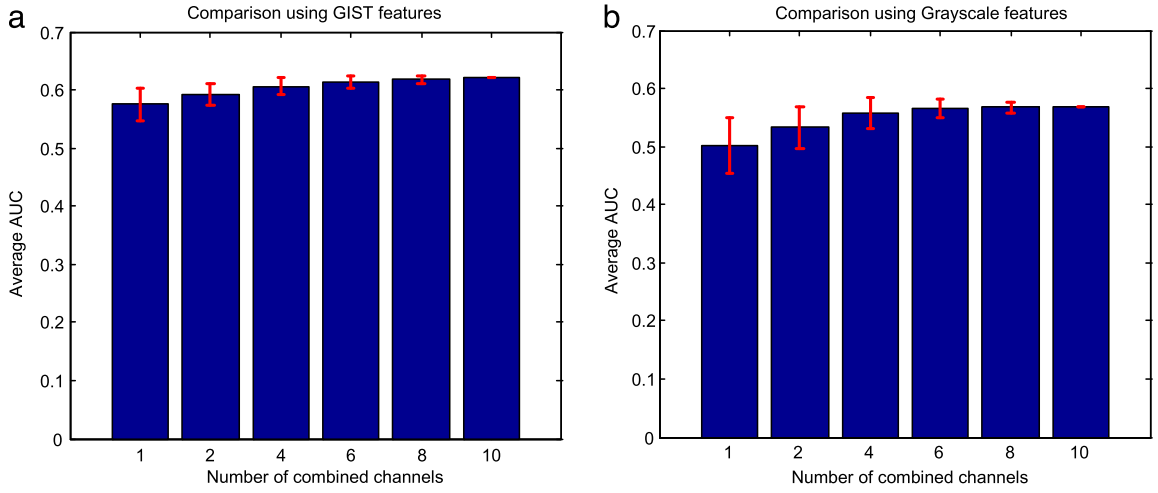

Fig. 18. Average AUC with errorbars for different numbers of combined channels on the St. Lucia dataset using GIST features (a) and Grayscale features (b). Number 2 in x axis indicates that the system combines any 2 channels from the total recognition channels.

5. Conclusions and future work

This paper has presented new research advancing the performance and capability of a biologically-inspired place recognition architecture based on the multi-scale grid cells found in the mammalian brain. In particular, it has proposed an *adaptive* approach, where the spatial scales are determined by the nature of the sensory input from the environment, rather than fixed beforehand. Automatic clustering methods are used to group the sensory input – in this case imagery – into natural groupings that

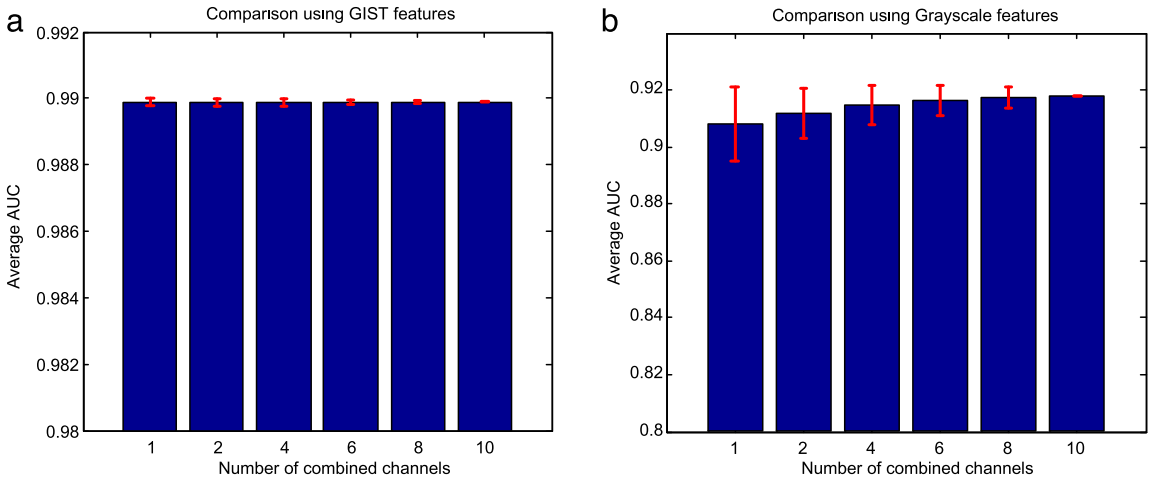

Fig. 19. Average AUC with errorbars for different numbers of combined channels on the Eynsham dataset using GIST features (a) and Grayscale features (b). Number 2 in *x* axis indicates that the system combines any 2 channels from the total recognition channels.

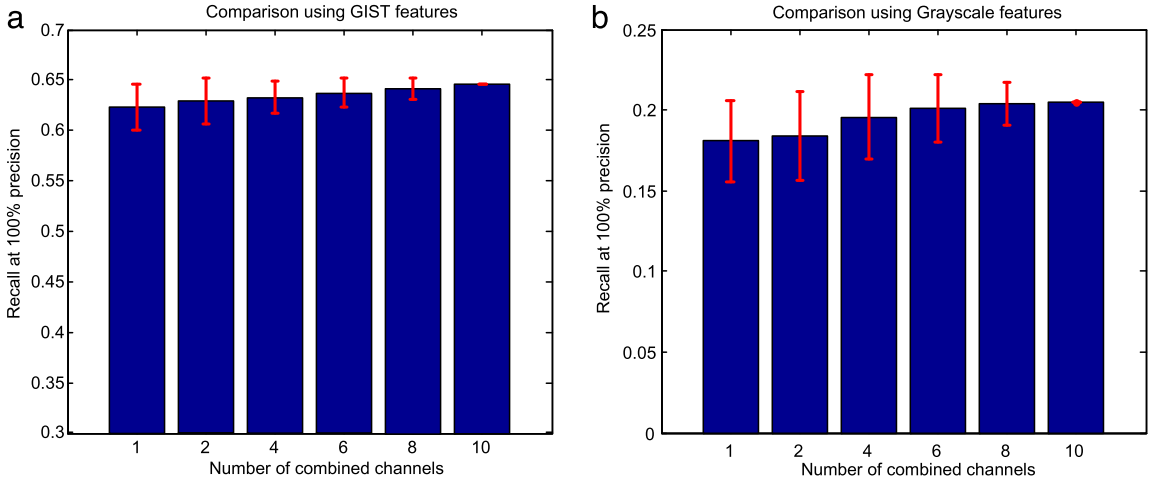

Fig. 20. Average maximal recall at 100% precision comparison with errorbars for different numbers of combined channels on the Eynsham dataset using GIST features (a) and Grayscale features (b). Number 2 in *x* axis indicates that the system combines any 2 channels from the total recognition channels.

then form the basis of each of the spatial scales being encoded. Experimental results on two benchmark datasets have demonstrated that the adaptive multi-scale place recognition approach leads to performance improvements over previous fixed multi-scale approach [3], and is also competitive with previous conventional place recognition techniques [9,10,18]. Performance increases with the number of combined channels but appears to plateau above 8 channels in Figs. 18–20. Neuroscience experiments have not yet revealed whether there is an upper limit for the number of channels—the results obtained in this research suggest from a computable performance basis in robot navigation that most of the performance gain is achieved by a modest number of place recognition channels.

Although our focus has been on the improvement offered by the adaptively multi-scale mapping system, it is apparent that the benefits are also dependent on the type of visual feature used. Perhaps not surprisingly, GIST features outperform Grayscale features in all experiments. This is reasonable because GIST features can capture higher semantic information and are therefore likely more discriminative. Future work will investigate the utility of other more powerful features such as those learned through training convolutional networks. The improvements from using an adaptive system shown in this paper are likely to hold for any feature-type; so as deep learning-based approaches improve, we expect to be able to apply this adaptive multi-scale method to whatever learnt features are available.

The current system can achieve significant performance improvements in one-dimensional path-like environments, in which the reference and testing datasets are obtained always along the same routes. This domain is applicable for robots and vehicles traversing path-like environments such as road networks. In addition, the current system assumes that the vehicle is moving at a constant speed while capturing images with a fixed frequency. In the future, we can extend the use of self-motion to aid the place recognition in an unconstrained movement, such as the indoor and aerial environment.

To make our work extendable to platforms navigating in open field environments, future work will extend the mechanism in our system to apply in two-dimensional environments. Preliminary non-directly bio-inspired work in this domain has suggested that

the conventional algorithms we have benchmarked against in this paper can be adapted to open-field environments [40]; we expect that the same will hold true for this adaptive multi-scale approach. The introduction of a motion calculation system for dead reckoning/path integration will also improve performance in a system operating in open areas. Absolute direction sensing in the form of a compass—such as provided by the regular skylight polarization pattern [41–43] will also improve performance in outdoor domains where such an addition is practicable. Rodents of course make use of multiple sensing modalities; it is likely that building on recent work performing sensor fusion using RatSLAM in [44,45] for this adaptive approach will yield further performance advances and new insights into how multiple sensing modalities contribute to multi-scale spatial encoding of space.

Acknowledgments

AJ and MM are with the Queensland University of Technology and MM is with the ARC Centre of Excellence for Robotic Vision. ZC is with ETH Zurich. This work was supported by funding from the National Nature Science Foundation of China under Grant 61573371 and 61503403, and an Asian Office of Aerospace Research and Development Grant FA2386-16-1-4027 and an Australian Research Council Future Fellowship FT140101229 to MM.

References

[1] B. Kuipers, Y.-T. Byun, A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations, Robot. Auton. Syst. 8 (1991) 47–63.
[2] M. Bosse, P. Newman, J. Leonard, M. Soika, W. Feiten, S. Teller, An Atlas framework for scalable mapping, in: Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on, 2003, pp. 1899–1906.
[3] Z. Chen, S. Lowry, A. Jacobson, M.E. Hasselmo, M. Milford, Bio-inspired homogeneous multi-scale place recognition, Neural Netw. 72 (2015) 48–61.
[4] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, E.I. Moser, Microstructure of a spatial map in the entorhinal cortex, Nature 436 (2005) 801–806.
[5] Y. Burak, I.R. Fiete, Accurate path integration in continuous attractor network models of grid cells, PLoS Comput. Biol. 5 (2009) e1000291.
[6] P.E. Welinder, Y. Burak, I.R. Fiete, Grid cells: the position code, neural network models of activity, and the problem of learning, Hippocampus 18 (2008) 1283–1300.
[7] B. Kuipers, J. Modayil, P. Beeson, M. MacMahon, F. Savelli, Local metrical and global topological maps in the hybrid spatial semantic hierarchy, in: Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on, 2004, pp. 4845–4851.
[8] Z. Chen, A. Jacobson, U.M. Erdem, M.E. Hasselmo, M. Milford, Multi-scale bio-inspired place recognition, in: 2014 IEEE international conference on robotics and automation, ICRA, 2014, pp. 1895–1901.
[9] M. Cummins, P. Newman, FAB-MAP: Probabilistic localization and mapping in the space of appearance, Int. J. Robot. Res. 27 (2008) 647–665.
[10] M. Cummins, P. Newman, Highly scalable appearance-only SLAM-FAB-MAP 2.0, in: Robotics: Science and Systems, 2009, pp. 12–18.
[11] A.J. Davison, I.D. Reid, N.D. Molton, O. Stasse, MonoSLAM: Real-time single camera SLAM, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 1052–1067.
[12] D. Ball, S. Heath, J. Wiles, G. Wyeth, P. Corke, M. Milford, OpenRatSLAM: an open source brain-based SLAM system, Auton. Robots 34 (2013) 149–176.
[13] M. Milford, G. Wyeth, Persistent navigation and mapping using a biologically inspired SLAM system, Int. J. Robot. Res. 29 (2010) 1131–1153.
[14] M.J. Milford, Robot Navigation from Nature.
[15] M.J. Milford, G.F. Wyeth, Mapping a suburb with a single camera using a biologically inspired SLAM system, IEEE Trans. Robot. 24 (2008) 1038–1053.
[16] K. Konolige, M. Agrawal, FrameSLAM: From bundle adjustment to real-time visual mapping, IEEE Trans. Robot. 24 (2008) 1066–1077.
[17] M. Milford, Vision-based place recognition: how low can you go? Int. J. Robot. Res. 32 (2013) 766–789.
[18] M.J. Milford, G.F. Wyeth, SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights, in: Robotics and Automation, ICRA, 2012 IEEE International Conference on, 2012, pp. 1643–1649.
[19] K. Konolige, E. Marder-Eppstein, B. Marthi, Navigation in hybrid metric-topological maps, in: Robotics and Automation, ICRA, 2011 IEEE International Conference on, 2011, pp. 3041–3047.
[20] S. Šegvić, A. Remazeilles, A. Diosi, F. Chaumette, A mapping and localization framework for scalable appearance-based navigation, Comput. Vis. Image Underst. 113 (2009) 172–187.
[21] J.C. Rangel, J. Martínez-Gómez, I. García-Varea, M. Cazorla, LexToMap: lexical-based topological mapping, Adv. Robot. 31 (2017) 268–281.
[22] H. Lee, R. Grosse, R. Ranganath, A.Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: Proceedings of the 26th annual international conference on machine learning, 2009, pp. 609–616.
[23] H. Stensola, T. Stensola, T. Solstad, K. Frøland, M.-B. Moser, E.I. Moser, The entorhinal grid map is discretized, Nature 492 (2012) 72–78.
[24] J. O'Keefe, J. Dostrovsky, The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat, Int. J. Comput. Vis. 34 (1971) 171–175.
[25] J. O'keefe, D. Conway, Hippocampal place units in the freely moving rat: why they fire where they fire, Exp. Brain Res. 31 (1978) 573–590.
[26] J.S. Taube, R.U. Muller, J.B. Ranck, Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis, J. Neurosci. 10 (1990) 420–435.
[27] N.J. Killian, M.J. Jutras, E.A. Buffalo, A map of visual space in the primate entorhinal cortex, Nature 491 (2012) 761–764.
[28] N. Ulanovsky, C.F. Moss, Hippocampal cellular and network activity in freely moving echolocating bats, Nature Neurosci. 10 (2007) 224–233.
[29] M.M. Yartsev, M.P. Witter, N. Ulanovsky, Grid cells without theta oscillations in the entorhinal cortex of bats, Nature 479 (2011) 103–107.
[30] I. Jolliffe, Principal Component Analysis, Wiley Online Library, 2002.
[31] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (2001) 145–175.
[32] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, 1973.
[33] S.L. Chiu, Fuzzy model identification based on cluster estimation, J. Intell. Fuzzy Syst. 2 (1994) 267–278.
[34] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Springer Science & Business Media, 2013.
[35] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (2005) 645–678.
[36] A.J. Glover, W.P. Maddern, M.J. Milford, G.F. Wyeth, FAB-MAP+ RatSLAM: Appearance-based SLAM for multiple times of day, in: 2010 IEEE International Conference on Robotics and Automation, ICRA, 2010 pp. 3507–3512.
[37] K.Q. Weinberger, J. Blitzer, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, in: Advances in Neural Information Processing Systems, 2005, pp. 1473–1480.
[38] M.M. Khan, D.R. Lester, L.A. Plana, A. Rast, X. Jin, E. Painkras, S.B. Furber, SpiNNaker: mapping neural networks onto a massively-parallel chip multiprocessor, in: Neural Networks, 2008. IJCNN 2008, IEEE World Congress on Computational Intelligence, IEEE International Joint Conference on, 2008, pp. 2849–2856.
[39] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 34 (2015) 1537–1557.
[40] J. Mount, M. Milford, 2D visual place recognition for domestic service robots at night, in: 2016 IEEE International Conference on Robotics and Automation, ICRA, 2016, pp. 4822–4829.
[41] C. Fan, X. Hu, J. Lian, L. Zhang, X. He, Design and calibration of a novel camera-based bio-inspired polarization navigation Sensor, IEEE Sens. J. 16 (2016) 3640–3648.
[42] D. Lambrinos, H. Kobayashi, R. Pfeifer, M. Maris, T. Labhart, R. Wehner, An autonomous agent navigating with a polarized light compass, Adapt. Behav. 6 (1997) 131–161.
[43] D. Lambrinos, R. Möller, T. Labhart, R. Pfeifer, R. Wehner, A mobile robot employing insect strategies for navigation, Robot. Auton. Syst. 30 (2000) 39–64.
[44] A. Jacobson, M. Milford, Towards brain-based sensor fusion for navigating robots, in: Proceedings of the 2012 Australasian Conference on Robotics & Automation, 2012.
[45] M. Milford, A. Jacobson, Brain-based sensor fusion for navigating robots, in: IEEE International Conference on Robotics and Automation, IEEE, Karlsruhe, Germany, 2013.

Chen Fan is a Ph.D. candidate in Robotics and Control in the School of mechatronics and Automation at National University of Defence Technology. He obtained his M.S. degree in Control Science and Engineering in 2013 from National University of Defence Technology. He also holds a B.S. degree in Electrical Engineering and Automation, awarded by Harbin Institute of Technology in 2011. His current research interests include vision-based place recognition and bio-inspired robotic navigation.

Zetao Chen is currently a Post-doctoral researcher in Swiss Federal Institute of Technology in Zurich (ETHZ). He received his Ph.D. degree in 2016 from Queensland University of Technology. Before that, he obtained his master degree in Artificial Intelligence in 2012 from University of Groningen and a bachelor degree in Electrical Engineering in 2009 from South China University of Technology. His current research interests include visual place recognition, bio-inspired robotics and visual-inertial navigation for small Unmanned Aerial Vehicles (UAVs).

Adam Jacobson is a Post-Doctoral Fellow at the Queensland University of Technology. He received his Bachelor of Engineering from Queensland University Technology and has submitted his Ph.D. in Biologically-Inspired Robot Localization and SLAM systems. His research involves the creation of biologically-inspired solutions for the deployment of multi-sensor SLAM navigation systems.

Xiaoping Hu is a Professor in the School of Mechatronics and Automation at National University of Defense Technology. He obtained the B.S. and M.S. degrees in automatic control systems and aircraft design from the School of Mechatronics and Automation at National University of Defence Technology in 1982 and 1985 respectively. His scientific interests include navigation, aircraft guidance and control and bio-inspired navigation.

Michael Milford (S'06–M'07) received the Ph.D. degree in electrical engineering and the Bachelor of Mechanical and Space Engineering degree from the University of Queensland, Brisbane, Australia. He is currently a Professor and Australian Research Council Future Fellow with the Queensland University of Technology (QUT), Brisbane, Australia, and a Chief Investigator for the Australian Centre of Excellence for Robotic Vision. He conducts interdisciplinary research into navigation and perception across the fields of robotics, neuroscience, and computer vision.