Neural Networks 72 (2015) 48–61

Contents lists available at ScienceDirect

Neural Networks

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

2015 Special Issue

Bio-inspired homogeneous multi-scale place recognition

Zetao Chen [a,b,*], Stephanie Lowry [a], Adam Jacobson [a,b], Michael E. Hasselmo [c], Michael Milford [a,b]

[a] School of Electrical Engineering and Computer Science, Queensland University of Technology, Australia

[b] Australian Centre for Robotic Vision, Queensland University of Technology, Australia

[c] Center for Memory and Brain and Graduate Program for Neuroscience, Boston University, United States

ARTICLE INFO

Article history:
Available online 29 October 2015

Keywords:
Bio-inspired
Multi-scale place recognition
Robot localization
Metric learning

ABSTRACT

Robotic mapping and localization systems typically operate at either one fixed spatial scale, or over two, combining a local metric map and a global topological map. In contrast, recent high profile discoveries in neuroscience have indicated that animals such as rodents navigate the world using multiple parallel maps, with each map encoding the world at a specific spatial scale. While a number of theoretical-only investigations have hypothesized several possible benefits of such a multi-scale mapping system, no one has comprehensively investigated the potential mapping and place recognition performance benefits for navigating robots in large real world environments, especially using more than two homogeneous map scales. In this paper we present a biologically-inspired multi-scale mapping system mimicking the rodent multi-scale map. Unlike hybrid metric-topological multi-scale robot mapping systems, this new system is homogeneous, distinguishable only by scale, like rodent neural maps. We present methods for training each network to learn and recognize places at a specific spatial scale, and techniques for combining the output from each of these parallel networks. This approach differs from traditional probabilistic robotic methods, where place recognition spatial specificity is passively driven by models of sensor uncertainty. Instead we intentionally create parallel learning systems that learn associations between sensory input and the environment at different spatial scales. We also conduct a systematic series of experiments and parameter studies that determine the effect on performance of using different neural map scaling ratios and different numbers of discrete map scales. The results demonstrate that a multi-scale approach universally improves place recognition performance and is capable of producing better than state of the art performance compared to existing robotic navigation algorithms. We analyze the results and discuss the implications with respect to several recent discoveries and theories regarding how multi-scale neural maps are learnt and used in the mammalian brain.

1. Introduction

The vast majority of robotic mapping and navigation systems perform mapping at either one fixed spatial scale or across two, typically comprising a local metric map and a topological global map. A range of recent high profile discoveries in neuroscience have demonstrated that animals such as rodents, and likely many other mammals including humans, encode the world using multiple parallel mapping systems, each of which encode the world at a different scale (Hafting, Fyhn, Molden, Moser, & Moser, 2005a, 2005b). In rodents, the mapping system scales from neurons that encode an area of a few square centimeters to neurons that encode an area of several square meters, with many intermediate scales represented in-between. Unlike hybrid metric-topological multi-scale robot mapping systems, rodent maps are homogeneous, distinguishable only by scale. While a number of theoretical-only investigations have hypothesized possible benefits of such a multi-scale mapping system (Burak & Fiete, 2009; Welinder, Burak, & Fiete, 2008), no one has comprehensively investigated the potential benefits of multi-scale mapping on place recognition in challenging real world environments.

In this paper, we present a biologically-inspired multi-scale mapping system mimicking the broad properties of the rodent multi-scale map. The first key innovation is to consider the place recognition problem as a hierarchical process—utilizing wider environmental context for more robust, coarser localization in

* Correspondence to: School of Electrical Engineering and Computer Science, Queensland University of Technology, 2 George St., Brisbane, QLD 4000, Australia. Tel.: +61 478891008.

E-mail address: zetao.chen@hdr.qut.edu.au (Z. Chen).

http://dx.doi.org/10.1016/j.neunet.2015.10.002
0893-6080/

parallel with finer localization on a smaller scale to improve localization accuracy. In this context, place recognition is framed not as the challenge of finding the single database image that best matches the current frame, but rather as one of finding all the database images within local spatial neighborhoods that are the best match for the sequence centered around the current frame. Our approach utilizes arrays of distance metrics, with each one trained to perform place recognition at a specific spatial scale, and a process for combining place recognition hypotheses from these different spatial scales. Unlike traditional probabilistic robotics methods, where spatial specificity is passively determined by sensor observation models, our approach intentionally creates parallel training systems to map the sensor input to the environment at different spatial scales.

This research extends on our previous work presented in Chen, Jacobson, Erdem, Hasselmo, and Milford (2013, 2014) in which we demonstrate that mapping over multiple scales uniformly improves place recognition performance over a single scale without sacrificing localization accuracy. We make three novel research contributions. Firstly, we introduce a metric learning-based algorithm to model the grid cells' discrete firing patterns. Secondly we propose an improved hierarchical framework to recognize places at multiple spatial scales. Lastly, for the first time our approach surpasses the performance of state of the art robotics algorithms, demonstrating the practical performance benefits of a homogeneous multi-scale mapping framework.

We conduct experiments on two robotics benchmark dataset and compare single- and multi-scale place recognition performance and demonstrate that multi-scale recognition leads to significantly improved recognition performance. We also conduct a systematic series of experiments and parameter studies that determine the effect on performance of using different neural map scaling ratios and different numbers of discrete map scales.

The paper is organized as follows. Section 2 discusses related place recognition and mapping techniques. In Section 3, we describe the components of the multi-scale place learning system. The experiments are detailed in Section 4, with results shown in Section 5. Finally we conclude the paper in Section 6 by discussing ongoing and future work.

2. Related work

Place recognition and mapping has been the subject of wide-ranging study both in the robotics and neuroscience community. This article is motivated by both fields, drawing inspiration from discoveries in neuroscience to develop novel multi-scale mapping algorithms for robots. To this end, we review the current state-of-the-art in place recognition algorithms for robots, including the existing use of multi-scale mapping within robotics. We briefly review evidence for multi-scale maps in the mammalian brain and note other bio-inspired mapping and navigation systems.

2.1. Place recognition methods

A fundamental challenge in mobile robotics is to develop robust navigation techniques. Place recognition – the ability to recognize places that the robot has already visited, and thereby correctly localize itself within the environment – is a key element of any navigation system. A great number of different sensors have been utilized for place recognition. Among them, visual sensors are the predominant sensor modality in many robot platforms with extensive research on vision-based place recognition (Angeli, Filliat, Doncieux, & Meyer, 2008; Cummins & Newman, 2008; Newman, Cole, & Ho, 2006; Ulrich & Nourbakhsh, 2000). The field of visual place recognition is well advanced, with place recognition systems being tested over paths measuring dozens (Schindler, Brown, & Szeliski, 2007) or even hundreds of kilometers (Cummins & Newman, 2009). Most appearance-based approaches start with image pre-processing (such as histogram normalization or noise removal), to improve image quality for future processing. Features are then extracted from the image, and a place matching process determines the most likely current position of the robot.

If multiple streams of data are available (such as multiple color channels) then a *voting* scheme (Ulrich & Nourbakhsh, 2000) can decide the robot location. Alternatively, a probabilistic calculation such as FAB-MAP (Cummins & Newman, 2008) can be used, where a likelihood model associating perception and location is learned on the extracted image features. FAB-MAP also compensates for *perceptual aliasing*; multiple locations may appear very similar and so observations must also be distinctive before FAB-MAP will match with high confidence. In RatSLAM (Milford, Wyeth, & Prasser, 2004), a biologically-inspired place recognition system based on a rat brain, localization is performed using a continuous attractor network (CAN) model combined with local view cells that excite and inhibit the elements in the neural network. Although RatSLAM is widely regarded as one of the state of the art biologically inspired robotic navigation systems, it is important to note that all of its benchmark achievements have come about due to a *single-scale* mapping system.

2.2. Multi-scale place recognition

In robotic navigation, multi-scale mapping often takes the form of a *hybrid metric-topological* or *topometric* map (Bosse et al., 2003; Konolige, Marder-Eppstein, & Marthi, 2011; Kuipers & Byun, 1991; Kuipers, Modayil, Beeson, MacMahon, & Savelli, 2004; Segvic, Remazeilles, Diosi, & Chaumette, 2009). Metric mapping develops geometrically accurate representations of the world, and allow centimeter-level accuracy in robot localization (Rowekamper et al., 2012), but is computationally infeasible over large areas, and struggles to close large loops (Bazeille & David, 2011). A compromise is to maintain small local metric submaps linked together in a topological map.

These mapping frameworks are *heterogeneous*, in that different types of maps (metric and topological) are used at different scales, and limited to two distinct scales. In contrast, in this research we consider *homogeneous* multi-scale mapping for robotics. This concept has been proposed (Kuipers, 1978, 2000) with topological *places* contained within a structure of topological *regions*. A similar concept to multi-scale topological mapping is the notion of summarizing an environment online, where the robot's observations are grouped into *topics* to allow for efficient summarization. This summarization can be performed using topic modeling (Paul, Rus, & Newman, 2012), coresets (Paul et al., 2012), Bayesian surprise (Girdhar & Dudek, 2012) or extremum summaries (Girdhar & Dudek, 2012). However, these environmental summarization techniques have not explicitly been used to perform place recognition (Theocharous, Murphy, & Kaelbling, 2004) investigate the concept of multi-scale robot localization and demonstrate that multiple scale representation helps to scale the H-POMDPs's algorithm to much larger models. However, up to now, there is still no quantitative evaluation on the benefits of multi-scale mapping in place recognition.

2.3. A multi-scale neuronal map

Over the past 30 years, there have been extensive studies on mapping and navigation mechanisms in rodents. Early studies focused on the part of the rodent brain known as the hippocampus, which was thought to be responsible for navigation tasks, and led to the discovery of *place cells* (O'Keefe & Dostrovsky, 1971) within the rat hippocampus which are only active when the animal is in

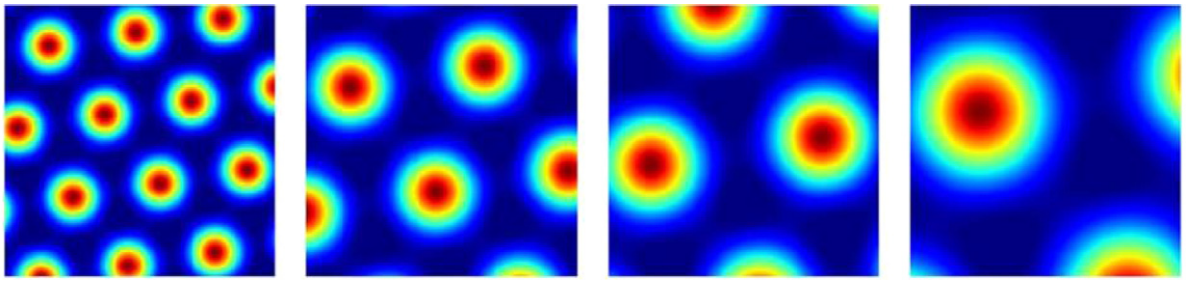

Fig. 1. Autocorrelograms of the firing fields of four simulated grid cells with varying scales. As grid firing fields become larger their relative spacing also increases, maintaining the same overall field structure.

a particular place in the environment. These place cells create an internal neural map of the environment (O'Keefe & Conway, 1978). Head direction cells are another type of cells discovered in the brain that fire whenever the animal is heading close to the cell's preferred angle (Taube, Muller, & Ranck, 1990). Later studies on the medial entorhinal cortex (MEC) led to the identification of *grid cells* (Hafting et al., 2005a, 2005b) in rats. Grid cells have also been found in the brains of monkeys (Killian, Jutras, & Buffalo, 2012) and bats (Ulanovsky & Moss, 2007; Yartsev, Witter, & Ulanovsky, 2011) and it is probable they also occur in humans (Doeller, Barry, & Burgess, 2010; Jacobs et al., 2013).

The area within which a particular grid cell fires is called the *place field* of that grid cell. Grid cells fire maximally whenever the animal is located at the vertices of a regular grid of equilateral triangles. Plotting the spatial autocorrelogram of the neural activity of a grid cell reveals the triangular tessellating nature of its firing field. Grid cells in the same area of the MEC fire with the same spacing and orientation, but with different phasing, and together cover every point of the environment. Furthermore, the whole population of grid cells encodes space at multiple scales (where the scale of a grid is defined as the distance between each place field). These discrete scales may occur in steps of approximately $\sqrt{2}$ (Stensola et al., 2012), although the value and consistency of this ratio is still to be determined conclusively. To simulate the neural activity of one grid cell, an electrode capable of recording neural activity of an individual cell was implanted into the brain of a rat to record the locations where that cell emits an action potential. An autocorrelogram of these firing locations is then plotted to better illustrate the neural activity of each grid cell. Fig. 1 shows the autocorrelograms of neural activities of four simulated grid cells, where the scales of the four 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 tens of square meters. The upper limit, if one indeed exists, is unknown. This integrated, multi-scale representation has been shown to have a number of theoretical advantages, including efficient mapping of arbitrarily large environments (Burak & Fiete, 2009; Welinder et al., 2008).

2.4. Neurally-inspired robotic mapping methods

The discoveries of place recognition and navigation models in rodent and human brains have inspired a number of place recognition models. Erdem and Hasselmo (Erdem & Hasselmo, 2014) propose a goal-directed navigation model which utilizes multiple types of simulated neurons (i.e. head direction cells, grid cells, place cells, reward cells and persistent spiking cells) to represent the environment. Experiments were conducted in a simulated arena with the environment being represented by place cells at different scales. Navigation in this model is based on a perfect self-motion sensor and has no other sensors, such as camera or ranger finder, to observe the external environment. This lack of perception components makes this model in its current

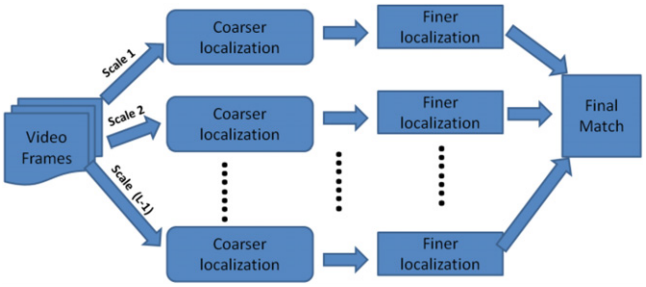

Fig. 2. Schematic of multi-scale place recognition. Each vertical row represents a different scale, each of which performs a coarser localization followed by a finer localization. The resulting multiple place recognition hypotheses are combined to produce a final match.

form inappropriate to perform with imperfect sensing because accumulated errors from self-motion sensor will rapidly violate the internal spatial representation. To partially address these practical shortcomings, the model was recently integrated with RatSLAM (Erdem, Milford, & Hasselmo, 2015).

Other models are designed more specifically for robotic applications and as such do incorporate some form of perception of the external environment, while using biologically-inspired concepts such as place cells (Giovannangeli, Gaussier, & Désilles, 2006; Milford et al., 2004). This paradigm has been shown to be remarkably effective with RatSLAM (Milford et al., 2004) achieving mapping of the longest path by a visual SLAM algorithm at the time (Milford & Wyeth, 2008), and a long term delivery robot experiment within a large office environment for two weeks (Milford & Wyeth, 2010).

These experimental results were achieved using a single-scale neural map. Using multiple networks with different mapping scales offers the potential to add a powerful additional combinatorial mechanism for improving place recognition performance and is the focus of the research presented in this paper.

3. Approach

The place recognition method presented here is inspired by the multi-scale grid cell structure of the rodent brain (Stensola et al., 2012). We describe the image features used, the machine learning methods used to model the grid cells and the mechanism of combining place recognition hypotheses from varying spatial scales to produce an overall place match hypothesis. A schematic of the place recognition process is shown in Fig. 2.

3.1. Feature extraction

The method makes no assumption about the feature types utilized. This paper evaluates three commonly used feature extraction methods—Grayscale Intensity 'Features' (pixels), Gist and deep learning features.

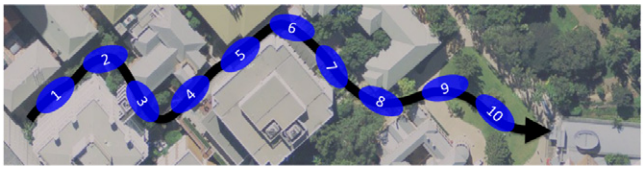

Fig. 3. Firing pattern example for one grid cell in a real dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1.1. Gray features

Grayscale images from the datasets were first down-sampled to 48×64 pixels and Principal Component Analysis (PCA) (Jolliffe, 2002) was applied. The top 500 principal eigenvectors were selected to use as the image feature as these were found to capture approximately 90% of the data variance.

3.1.2. Gist features

A variety of experimental studies have demonstrated that humans perform rapid categorization of a scene by integrating only the coarse global information or "gist" (Biederman, 1988; Potter, 1975). Using the model proposed by Oliva and Torralba (Oliva & Torralba, 2001), we extracted Gist features from down-sampled 48×64 images resulting in a 512-dimensional feature. Once again, PCA was applied. In this case the top 300 principal eigenvectors were found to capture approximately 90% of the total variance and were thus used as the image features.

3.1.3. Deep learning features

We use the output of the final convolution layer from a pretrained network called Overfeat (Sermanet et al., 2013) as the deep learning features. The Overfeat network is trained on the ImageNet 2012 dataset (Deng et al., 2009), which consists of 1.2 million images and 1000 classes. The original 3072-dimensional vector was reduced by PCA to 1500 dimensions which capture about 90% data variance.

3.2. A learning algorithm for modeling grid cells

To model the overlapping discrete firing patterns of grid cells Stensola, (Stensola et al., 2012), our place recognition system learns a pattern of grid cells across the environment. Fig. 3 demonstrates neural activity of one such grid cell in a real dataset. The black arrow indicates the route the system is trained to recognize, and the blue circles represent firing clusters of that grid cell. The size of the blue circle represents the spatial size of the grid cell, and the video frames that are captured within a particular blue circle are all assigned the same label (as indicated by the number on top of the blue circle). All frames that are captured outside any blue cluster are considered as the vanishing area of the grid cell and are assigned to another grid cell.

3.3. Modeling single-scale grid cell firing

The system uses Large Margin Nearest Neighbors (LMNN) (Weinberger, Blitzer, & Saul, 2005) to learn arrays of distance metrics. Each array maps the data to a new space where places can be recognized at a specific spatial scale. All the frames from the data, along with their labels (numbers on blue circles in Fig. 3), are used by the LMNN process to train a distance metric to map the data to a space such that the k nearest neighbors of any frame inside a cluster always come from the same cluster.

Given the feature vectors extracted from the images, LMNN learns a Mahalanobis distance metric which clusters images from spatially approximate regions in a way that mimics grid cell behavior. Training data can be denoted as:

$$\{x_i, y_i\}_{i=1}^{n} \in R^N \times \{1, 2, \ldots, C\} \tag{1}$$

where x_i denotes the feature vector in N dimensional space and y_i is the label with C different classes.

The LMNN training procedure consists of two steps. The first step involves identifying a number of k nearest neighbors for each input x_i. In this paper, we select the target neighbors by simply computing k nearest neighbors from frames in the same cluster using the Euclidean distance. The notation $j \rightarrow i$ indicates instance x_j is a target neighbor of instance x_i.

The second step is to train a Mahalanobis distance metric M such that all target neighbors x_j are closer to x_i than any other samples with different labels. A Mahalanobis distance metric M computes the distance between x_i and x_j as:

$$d_M\left(x_i, x_j\right) = \left(x_i - x_j\right)^T M \left(x_i - x_j\right) \tag{2}$$

where $M \succcurlyeq 0$ should be positive semidefinite to generate a positive distance measurement. When M is an identity matrix, Eq. (2) is reduced to the Euclidean distance metric. A positive definite matrix can be achieved by constructing:

$$M = \beta^T \beta \tag{3}$$

where β is an arbitrary real $N_1 \times N_2$ matrix with $N_2 \leq N_1$. In this paper we only consider the case $N_2 = N_1$. Substituting Eq. (3) into Eq. (2) results in:

$$d_M\left(x_i, x_j\right) = \left(x_i - x_j\right)^T \beta^T \beta \left(x_i - x_j\right) = \left\| \beta \left(x_i - x_j\right) \right\|^2 \tag{4}$$

where the matrix β maps the data to a new space in which Euclidean distance is calculated.

We let $y_{ij} \in \{1, 0\}$ denote whether or not x_i and x_j share the same label and $\varepsilon_{ijl} \geq 0$ indicate the amount by which a differently labeled sample x_l invades the boundary of x_i defined by all its neighbors $\sum_{j \rightarrow i} x_j$. The metric M is computed by solving the following semidefinite program:

$$\text{Minimize} \sum_{j \rightarrow i} \left[d_M\left(x_i, x_j\right) + \mu \sum_l \left(1 - y_{il}\right) \varepsilon_{ijl} \right] \tag{5}$$

subject to:

(a) $d_M\left(x_i, x_l\right) - d_M\left(x_i, x_j\right) \geq 1 - \varepsilon_{ijl}$

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

(c) $M \succeq 0$.

The constant μ controls the trade-off between the two terms in the objective function and is set using cross-validation. The constraint (a) penalizes when any differently labeled input x_l invades the local neighbors of x_i. Such an invader generates a positive slack variable ε_{ijl}. The constraint (b) enforces a positive value of slack variable ε_{ijl}, and the constraint (c) enforces metric M to be positive definite. Since the distance $d_M(x_i, x_j)$ is linear in the matrix M, the above optimization is a semidefinite program and a global optimum can be efficiently computed.

3.4. Modeling multiple grid cells

Recordings of multiple grid cells show that grid cells encode multiple, discrete scales of place fields, but that grid cells with similar spatial scales can fire in an overlapping way (Stensola et al., 2012). Fig. 4 illustrates such a situation with an example of eight grid cells that model two different spatial scales. The first group of four cells fires in a more specific scale (S_1) and the second four grid cells fire in a coarser scale (S_2). Within each group, the cells fire in regions that overlap.

Algorithm 1 Pseudocode for single-scale place recognition (Section 3.5)

0: Initialization:
  Spatial scale: S_c
  Number of distance metrics at scale: S_c
  Testing cluster P_j in scale S_c
1: For each distance metric at scale S_c:
2: Search for the best matching training sequence Y using Equation (7)
3: Pick the final hypothesis $P(S_c)$ for P_j at scale S_c using equation (8)
4: Return the final hypothesis $P(S_c)$ as the training sequence that best matches P_j

Algorithm 2 Pseudocode for multi-scale verification (Section 3.6)

0: Initialization:
  All L spatial scale: S_c $(c = 1,2,\ldots,L)$ where $S_1 = S_{N_{min}}$ and S_L is the largest scale.
  Testing cluster P_j in scale $S_{N_{min}}$
1: For each coarser spatial scale $S_c (c = 2,3,\ldots,L)$ (Section 3.6.1)
    Pick a testing cluster of scale S_c centered around P_j
    $P(S_c)$ ← Coarser localization at S_c using Algorithm 1
    $P(S_{c,N_{min}})$ ← Finer localization at scale $S_{N_{min}}$ at space $[P(S_c)\ (P(S_c) + (S_c) - 1)]$ using Algorithm 1
2: Return the best match P(final) using Equation (9) (Section 3.6.2)

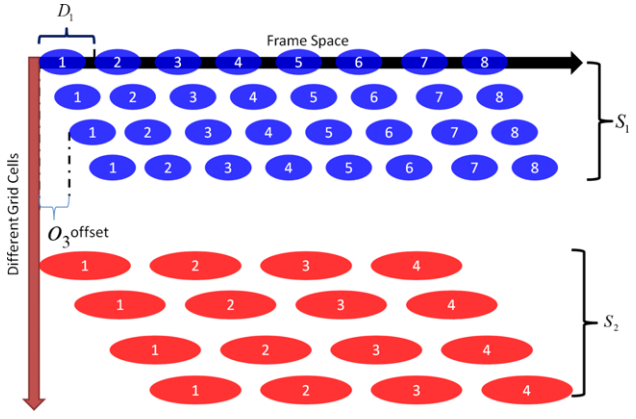

Fig. 4. Firing pattern of eight grid cells in two different spatial scales (red and blue) in overlapping pattern. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We denote O_t as the offset between the first frame in the dataset and the first firing frame in each cell. Assuming we have L different spatial scales to model: S_k $(k = 1,\ldots,L)$ and for each spatial scale S_k, the distance between consecutive initial firing points is $D_k (k = 1,\ldots,L)$ (Fig. 4). For each spatial scale S_k, we model the overlapping firing pattern by training N different Mahalanobis distance metrics with each one modeling the discrete firing pattern in scale S_k with different initial offsets

$$O_m = \frac{D_k\,(m-1)}{N}, \quad m = 1,\ldots,N. \tag{6}$$

We denote the distance metrics after metric learning as $M_{k,m} (k = 1,\ldots,L,\ m = 1,\ldots,N)$.

3.5. Single-scale place recognition using multiple grid cell models

This section describes how to recognize places at a particular spatial scale using distance metric trained from the previous section. Testing image frames are first grouped temporally into consecutive clusters (Fig. 4). The size of each consecutive cluster determines the accuracy that the localization system can achieve. A cluster with smaller size will have higher localization accuracy. Assume the current testing cluster size is S_c and images in that cluster are denoted as $P_j (j = 1,\ldots,S_c)$. Training images are denoted as $T_i (i = 1,\ldots,T)$ and the image difference between T_i and P_j using distance metric M is denoted as $D_M(i,j)$. For each testing cluster, localized image sequence matching is performed through the whole training space (Fig. 5) to search for the particular grid place field Y that best matches the testing cluster:

$$Y = \arg\min_{p} \sum_{i=p}^{i=p+S_c-1} D_M(i, i-p+1), \quad \forall p \in [1, T-S_c+1] \tag{7}$$

with corresponding firing score: $F(Y) = \sum_{i=Y}^{i=Y+S_c-1} D_M(i, i-Y+1)$. A smaller value of $F(Y)$ indicates a stronger and more confident match.

Since there are N different metrics trained in each spatial scale S_c, N different place recognition hypotheses will be produced: $Y_{c,m}$, $(m = 1,\ldots,N)$ with corresponding firing scores $F\left(Y_{c,m}\right)$, $(m = 1,\ldots,N)$. The final hypothesis $P(S_c)$ reported at scale S_c is the one with the smallest firing score:

$$P(S_c) = \arg\min_{m} F\left(Y_{c,m}\right), \quad m = 1,\ldots,N. \tag{8}$$

The pseudocode for single-scale place recognition is listed in Algorithm 1.

3.6. Multi-scale place matching verification

A localization system from a spatial scale is capable of producing place recognition hypotheses that are only as precise as the average size of a segment for that scale. Thus a system with a cluster size of 6 frames will report place recognition hypotheses that are twice as spatially specific as a system with 12 frames. Here we present a two-step method for combining place recognition hypotheses at multiple spatial scales (see Fig. 6). The pseudocode for multi-scale verification is illustrated in Algorithm 2.

3.6.1. Coarser-to-finer localization

The system performs coarse localization at each of the larger spatial scales followed by finer localization at the smallest spatial scale to provide a more accurate estimation. Assuming there are L different spatial scales to combine, the smallest spatial scale $S_{N_{min}}$ is used for finer localization and each of the other $(L-1)$ larger scales is used for coarser estimation. Each of the $(L-1)$ coarser scales generates a coarse estimation at space $[P(S_c)\ (P(S_c) + (S_c) - 1)]$ within which a finer search at scale $S_{N_{min}}$ is used to produce a hypothesis at scale $S_{N_{min}}$. A place recognition hypothesis at the smallest spatial scale $S_{N_{min}}$ produced from a coarser estimation at scale S_c is denoted as: $P\left(S_{C,N_{min}}\right)$, $C = (1,\ldots,(L-1))$ with corresponding firing score $F(S_{C,N_{min}})$.

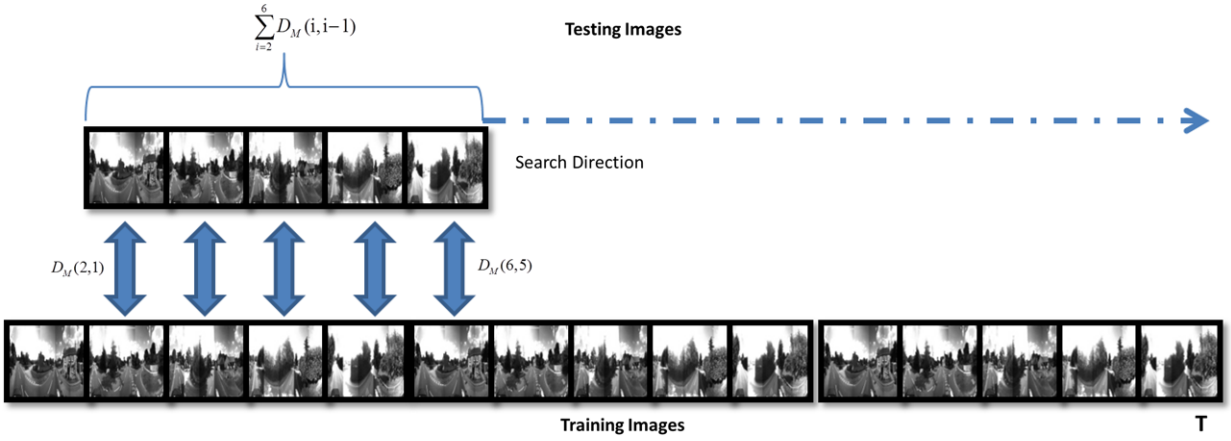

Fig. 5. Searching for a sequence matching a five-frame test sequence.

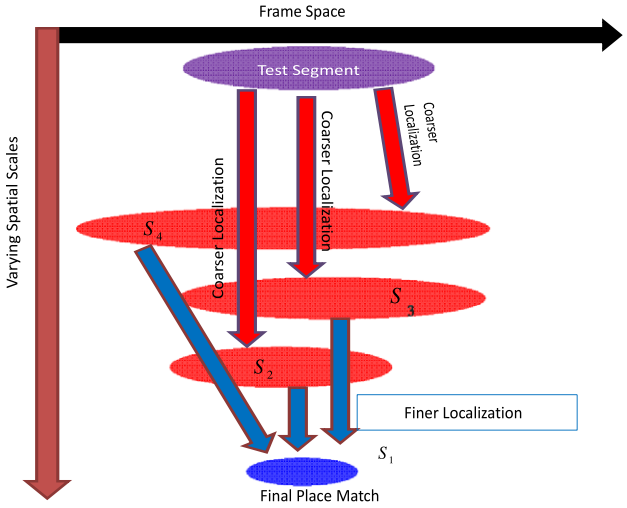

Fig. 6. A two-step method to combine hypotheses at multiple scales. Each of the three coarser scale ($\mathbf{S_2}$, $\mathbf{S_3}$, $\mathbf{S_4}$) provides a coarser estimation, followed by a finer localization at the smallest scale at $\mathbf{S_1}$. The hypothesis with smallest firing score at the smallest scale is picked as the final match.

3.6.2. Finding the best match candidate

At the smallest spatial scale, the $(N-1)$ competing place recognition hypotheses can be different. To determine the most likely hypothesis, we pick the one with the lowest firing scores:

$$P(final) = \arg\min_{c} F(S_c), \quad c = 1, \ldots, (N-1). \tag{9}$$

4. Experiments

In this section, we describe the dataset used and the training and testing procedures.

4.1. Datasets

The first dataset comprises camera data from a car traveling along a selection of streets in the suburb of St. Lucia, Brisbane, first presented in Glover, Maddern, Milford, and Wyeth (2010). The video was captured using a web camera at 640 × 480 pixel resolution at an average frame rate of 15 frames per second. An aerial overhead map was shown in Fig. 7(a). The route was traversed once in the morning and then once more in the afternoon. We picked a subset of about 2000 images for experiment. GPS data was logged at 1 Hz for ground truth. We simulated the odometry information by linearly interpolating the GPS information and used the odometry information to drive the cluster formulation as discussed in Section 4.2.1.

Table 1
Spatial scales for testing in Eynsham dataset.

Scaling ratio	Scale one	Scale two	Scale three	Scale four	Scale five	Scale six
$2^{1/4}$	6	8	9	11	12	15
$2^{2/4}$	6	9	13	17	25	34
$2^{3/4}$	6	11	17	29	49	81
$2^{4/4}$	6	12	24	48	96	192
$2^{5/4}$	6	15	34	81	192	457

The second dataset was the Eynsham dataset (Fig. 7(b)) which is a large 70 km road-based dataset (2 × 35 km traverses) used in the FAB-MAP (Cummins & Newman, 2009) and SeqSLAM studies (Milford, 2013; Milford & Wyeth, 2012). Panoramic images were captured at 7 m intervals using a Ladybug 2 camera. The dataset consists of two traverses along the same route. The dataset provides GPS-derived ground truth. In this experiment, images located within 40 m of each other were deemed to be correct matches, consistent with the tolerance used in the original study (Cummins & Newman, 2009).

4.2. Training and testing procedure

Images from the first traverse of the environment were used for training while images from the second traverse were used to evaluate performance. The overall training procedure consisted of the following three steps: dataset segmentation, feature extraction and metric learning.

4.2.1. Dataset segmentation

We will refer to the spatial scale $S_k (k = 1, \ldots, L)$ by the number of frames in each firing cluster. For example, a spatial scale of 34 means that each of the firing clusters in that scale contains 34 frames. For the Eynsham dataset, we chose the smallest spatial scale to be 6 frames, which corresponds to about 40 m in accuracy. This scale is consistent with the 40 m ground-truth tolerance used in the original study (Cummins & Newman, 2009). Because the Eynsham frame captures were triggered at regular distance intervals by a GPS, for this specific dataset frame number and distance are analogous. The coarser scales used scaling ratios of $2^{1/4}$, $2^{1/2}$, $2^{3/4}$, $2^{4/4}$, $2^{5/4}$ (see Table 1). For the St. Lucia dataset, we

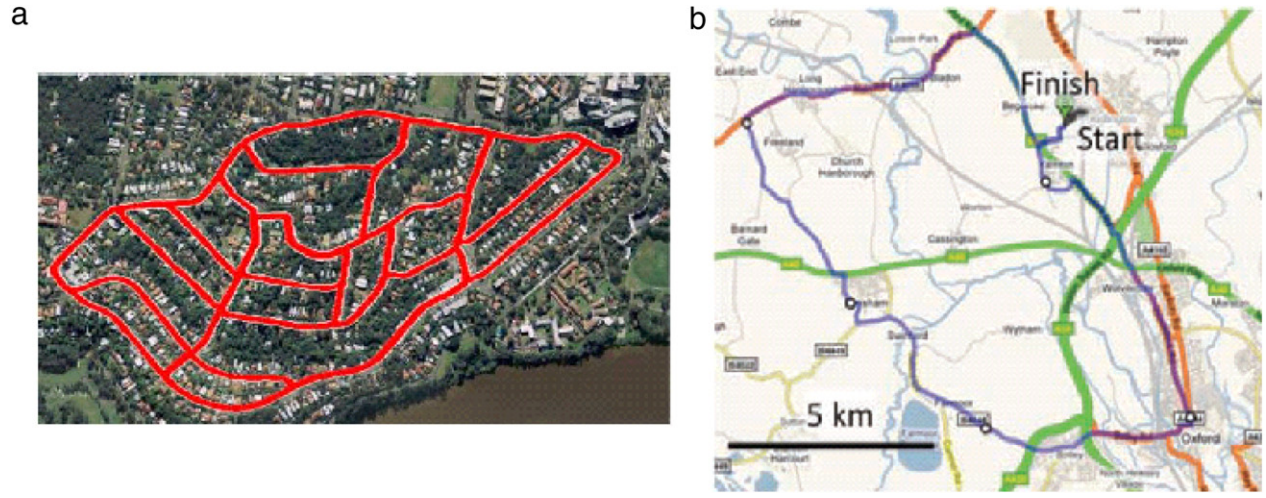

Fig. 7. Aerial overhead images showing the route of the (a) St. Lucia and (b) Eynsham dataset. (Imagery ©2012 Cnes/Spot Image, DigitalGlobe, GeoEye, Sinclair Knight Merz & Fugro.)

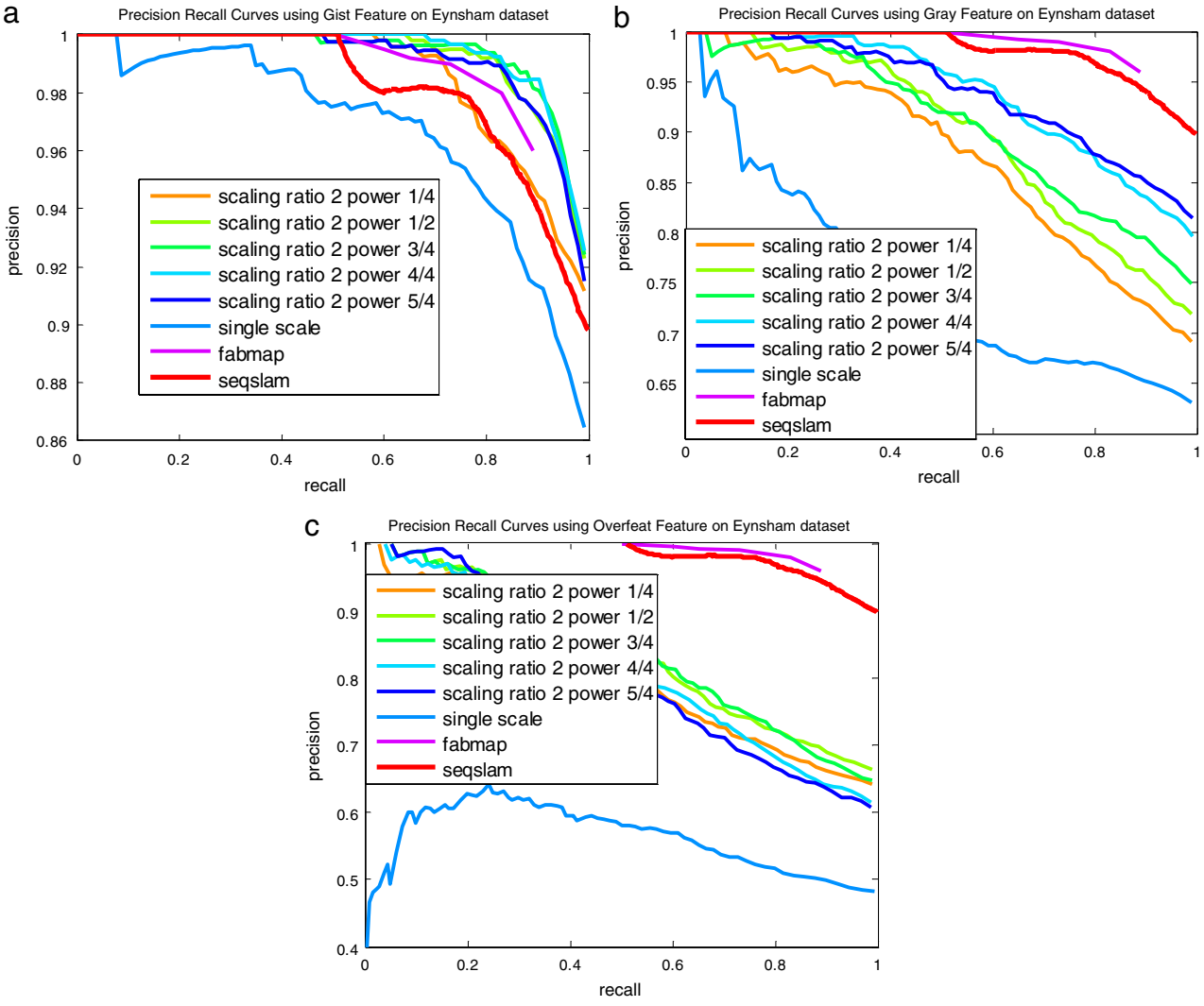

Fig. 8. Precision–recall curves demonstrating the single- (“6 frames system”) and multi-scale (“combined different scaling ratio”) place recognition performance using gist feature (a), gray feature (b) and deep learning features (c) on the Eynsham dataset.

utilized the estimated odometry information to drive the cluster formation, and chose the smallest spatial scale to be 2 m which corresponds to about 4 frames.

4.2.2. Feature extraction

Feature extraction enabled us to accelerate the training and testing stages by providing a reduced dimensionality input into

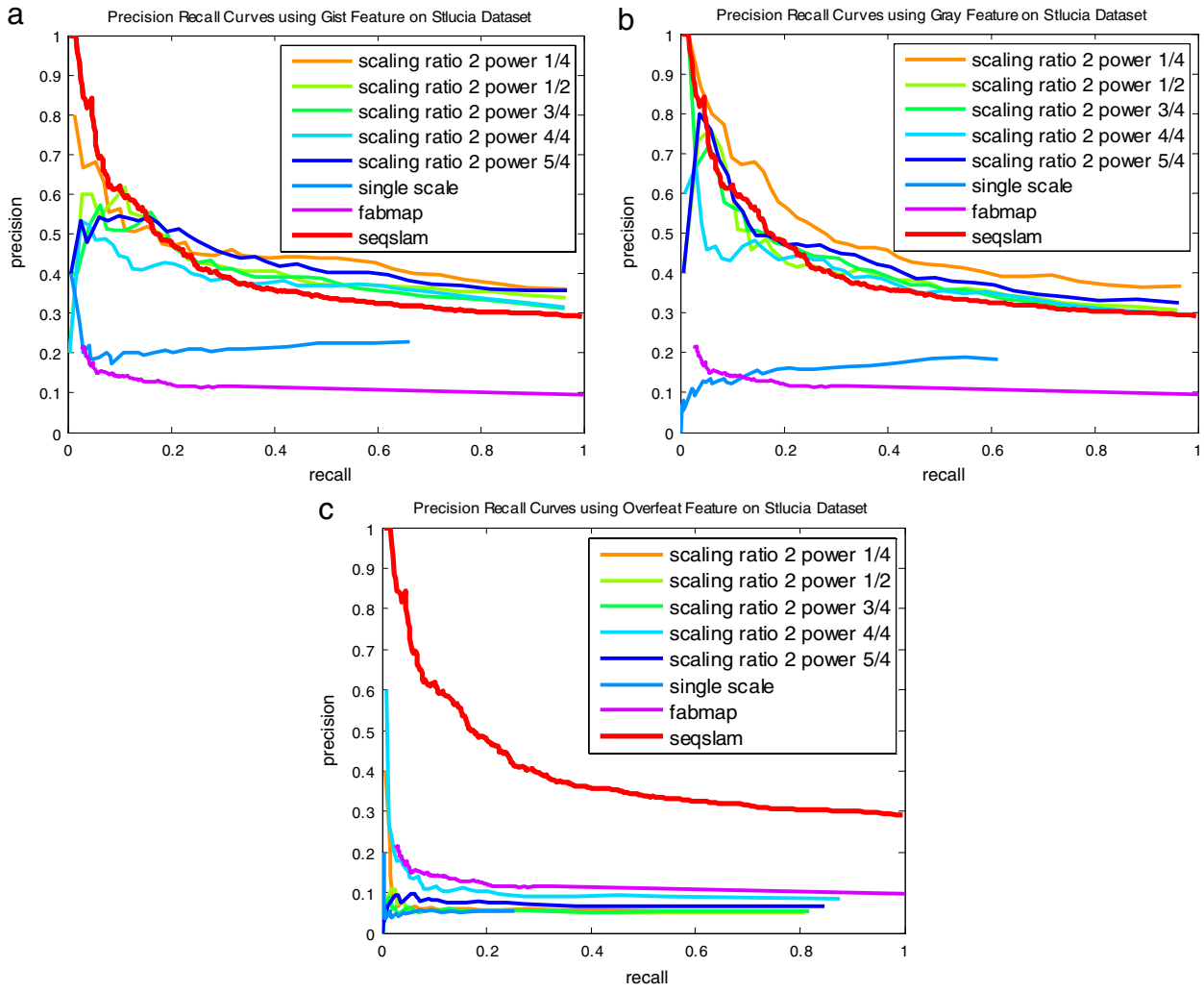

Fig. 9. Precision–recall curves demonstrating the single- ("2 meters system") and multi-scale ("combined different scaling ratio") place recognition performance using gist feature (a), gray feature (b) and deep learning features (c) on the St.Lucia dataset.

the LMNN models. Three feature types (as discussed in Section 3.1) were extracted: Gray, Gist and Overfeat.

4.2.3. Metric learning

The third step involved training a distance metric for each spatial scale. The first traverse of the route was used for training and the second traverse was used for testing. During the training, 20% of the training data was used as validation set to prevent over-fitting and determine early stopping.

5. Results

We present a performance comparison between single, multi-scale place recognition and one state-of-the-art algorithm—FABMAP (Cummins & Newman, 2009), a parameter study to determine the effect on performance of using different neural map scaling ratios and different numbers of discrete map scales, and an illustrative multi-scale place recognition combination plot.

5.1. Single- and multi-scale place recognition

In a place recognition task, 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 correct places in the dataset. A common strategy to combine both measurements is to use a precision–recall curve which is a two-dimensional plot with the x-axis indicating recall and the y-axis describing precision. Perfect performance occurs when the precision remains equal to one, as the recall increases from 0 to 1.

Figs. 8 and 9 present precision–recall (PR) curves resulting for the SeqSLAM, FABMAP, the single- and multi-scale place recognition experiments using Gist, Gray and deep learning features on the Eynsham dataset (Fig. 8) and the St. Lucia dataset (Fig. 9).

Five set of combinations are shown—"combined scaling ratio $2^{1/4}$, $2^{2/4}$, $2^{3/4}$, $2^{4/4}$ and $2^{5/4}$ respectively", as well as results from using a single scale, SeqSLAM and FABMAP algorithm. The results show that multi-scale matching consistently improves the performance. On the Eynsham dataset, using Gist features improves the recall rate at 100% precision by a factor of nearly 8 from 8% using a single-scale to about 68% when using the scaling scale of $2^{4/4}$. The corresponding best improvement using Gray features is from 3% to about 22%. Using the Overfeat feature, the maximal recall at 100% precision can be improved from 0% to about 5% by using a scaling ratio of $2^{5/4}$. A scaling ratio of $2^{4/4}$ achieves the best performance both on the Gist feature and PCA features. On the Overfeat feature, different scaling

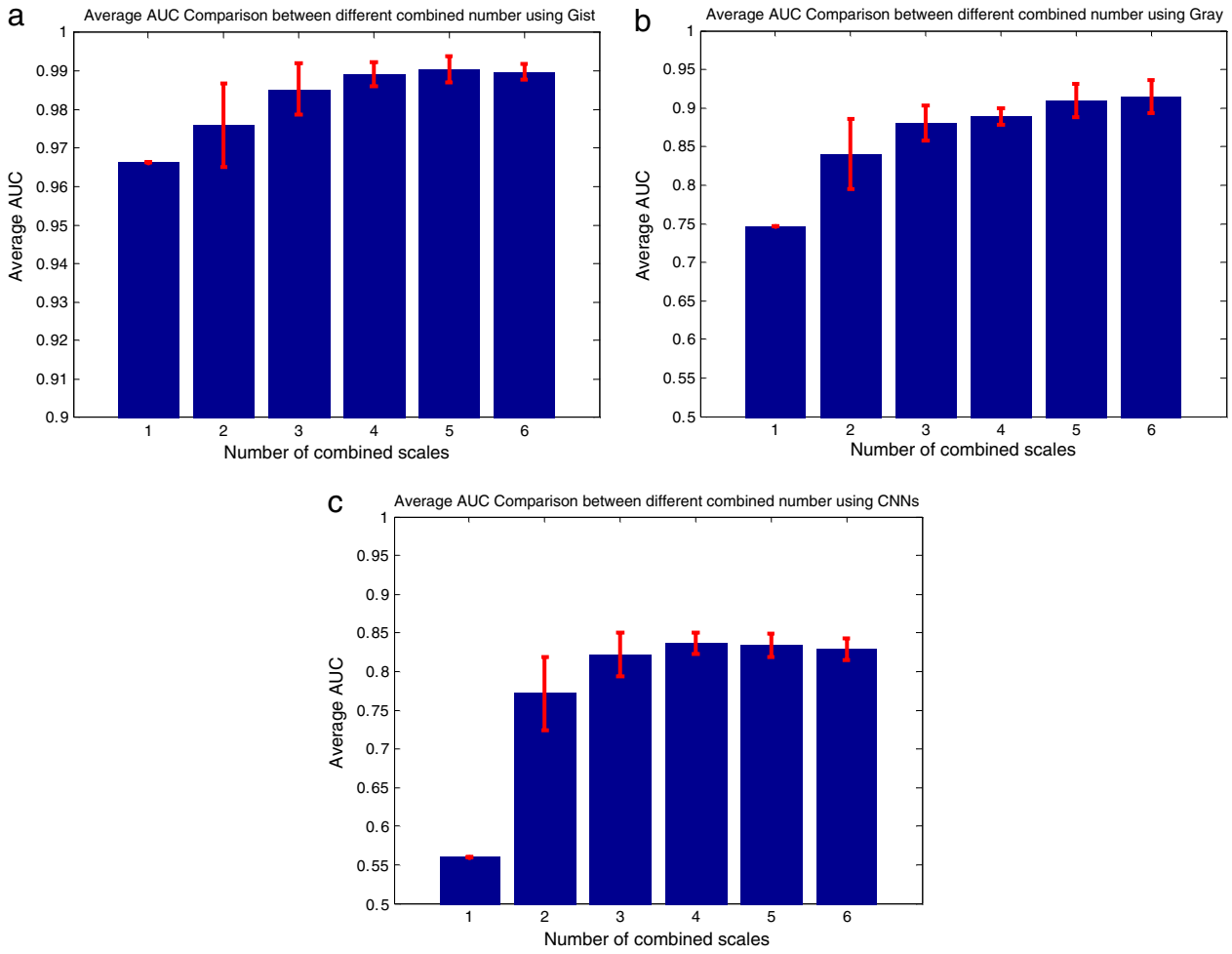

Fig. 10. Average AUC with errorbars for different numbers of combined scales on the Eynsham dataset using Gist features (a), Gray features (b) and deep learning features (c). Number 1 in the *x* axis indicates the performance of single-scale matching.

ratios deliver varying performance. On the St. Lucia dataset, the same improvement can be observed when multiple scales are combined to perform place recognition. On the Gist and Gray features, all different scaling ratios outperform the state-of-the-art FABMAP algorithm, indicating that this multi-scale algorithm may demonstrate a more clear advantage in appearance-changing environments. On the Gist feature, different scaling ratios deliver relatively similar performance, while on the Gray features, a scaling ratio of $2^{1/4}$ achieves a slightly higher improvement than others. On the Overfeat feature, the improvement introduced by multi-scale system is not very obvious, possibly because that features extracted at the top layer of a deep network are not very discriminative in a condition-changing environment (Chen, Obadiah, Jacobson, & Milford, 2014).

Although our focus is on the improvement potential offered by adopting a multi-scale approach, we provide absolute comparison metrics here as well. The best performance on the Eynsham dataset is achieved using the Gist feature. The maximum recall rate of about 68% at 100% precision is superior to the state-of-the-art 51% recall rate achieved by SeqSLAM (Milford, 2013) and 49% recall rate achieved by FABMAP (Cummins & Newman, 2009). On the St. Lucia dataset, when using the Gist or Gray features, the multi-scale system can consistently outperform the state-of-the-art FABMAP and SeqSLAM algorithm, in most of the scaling ratios utilized.

5.2. Systematic parameter studies

In this section, we conduct a systematic series of experiments and parameter studies on the effect of performance using different neural map scaling ratios and different numbers of discrete map scales. The performance on the Eynsham dataset is evaluated by both the Area Under the Curve (AUC) on the precision–recall curve (Figs. 10 and 13) and the maximal recall at 100% precision (Figs. 11 and 14). 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. The St. Lucia dataset is evaluated by only the AUC (Figs. 12 and 15), because all scaling ratios cannot achieve 100% precision, making maximal recall at 100% precision always 0%. We also conduct and report some important statistical tests of significance at 95% confidence level between using different experiment parameters.

Figs. 10 and 11 evaluate the influence of the number of combined scales on the performance of the Eynsham dataset, evaluated by the average AUC (Fig. 10) and maximal recall at 100% precision (Fig. 11) over all scale ratios. Using the Gist and Gray features, more scales always delivers better AUC performance, while using the Overfeat feature, combining different scales achieves statistically similar performance improvement over using a single scale. The maximal recall demonstrates a slightly different tendency: with both the Gist and Gray features, the best performance is achieved when combining five spatial scales and

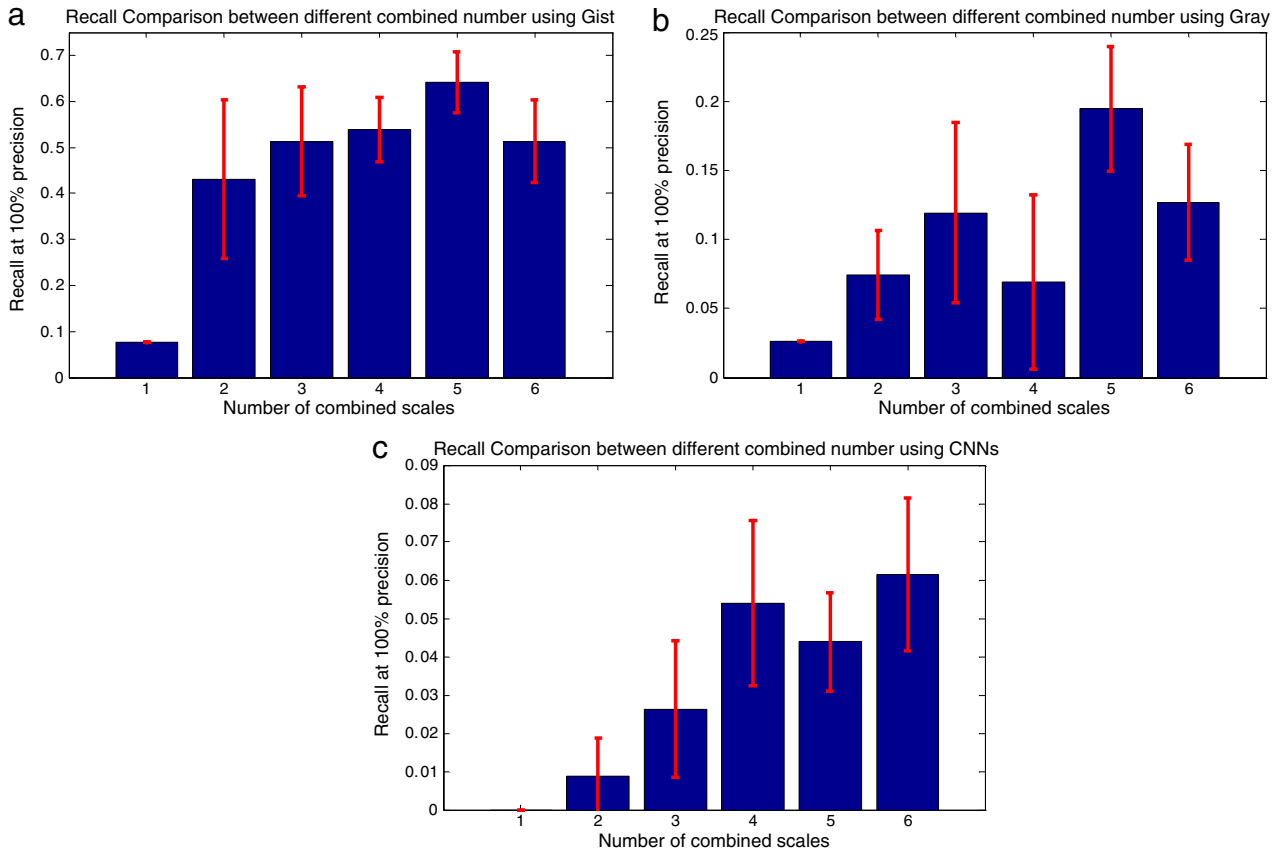

Fig. 11. Maximal recall at 100% precision comparison with errorbars for different numbers of combined scales in the Eynsham using Gist features (a), Gray features (b) and deep learning features (c). Number 1 in the *x* axis indicates the performance of single-scale matching.

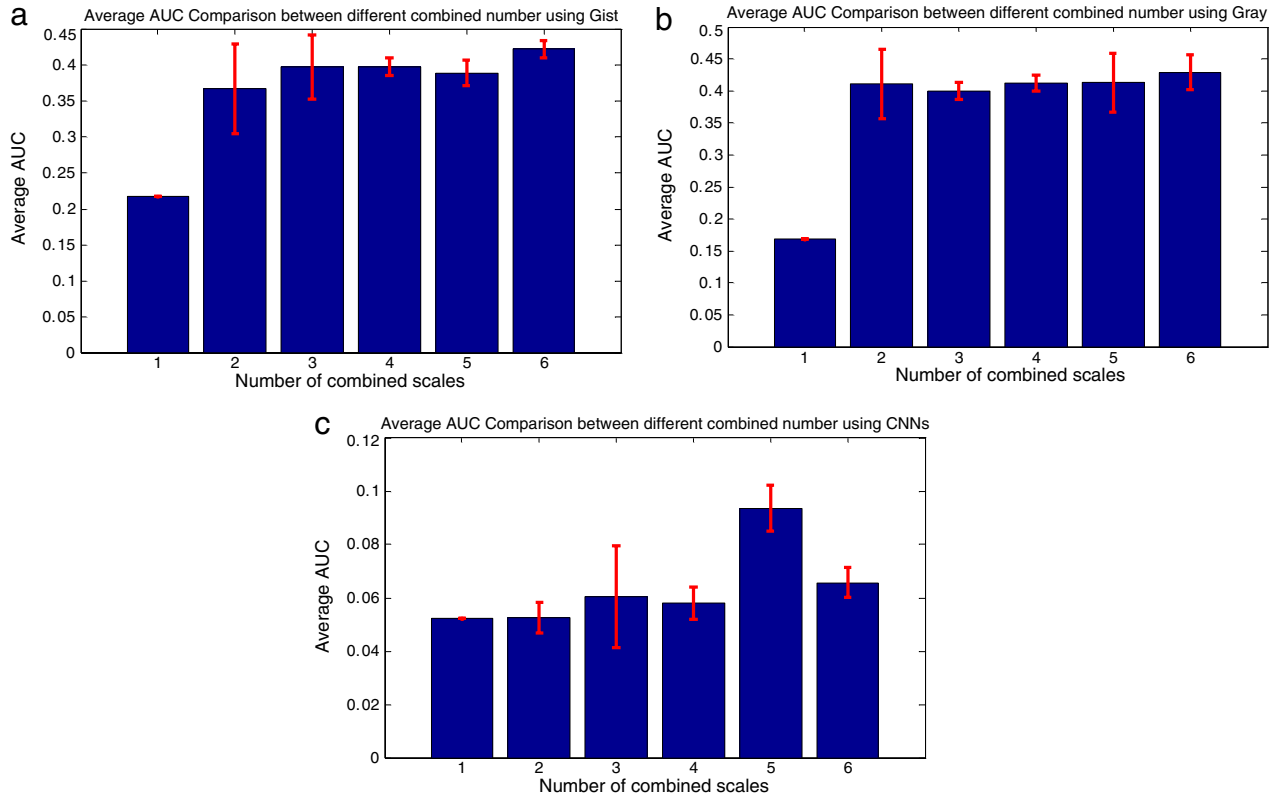

Fig. 12. Average AUC with errorbars for different numbers of combined scales on the St. Lucia dataset using Gist features (a), Gray features (b) and deep learning features (c). Number 1 in the *x* axis indicates the performance of single-scale matching.

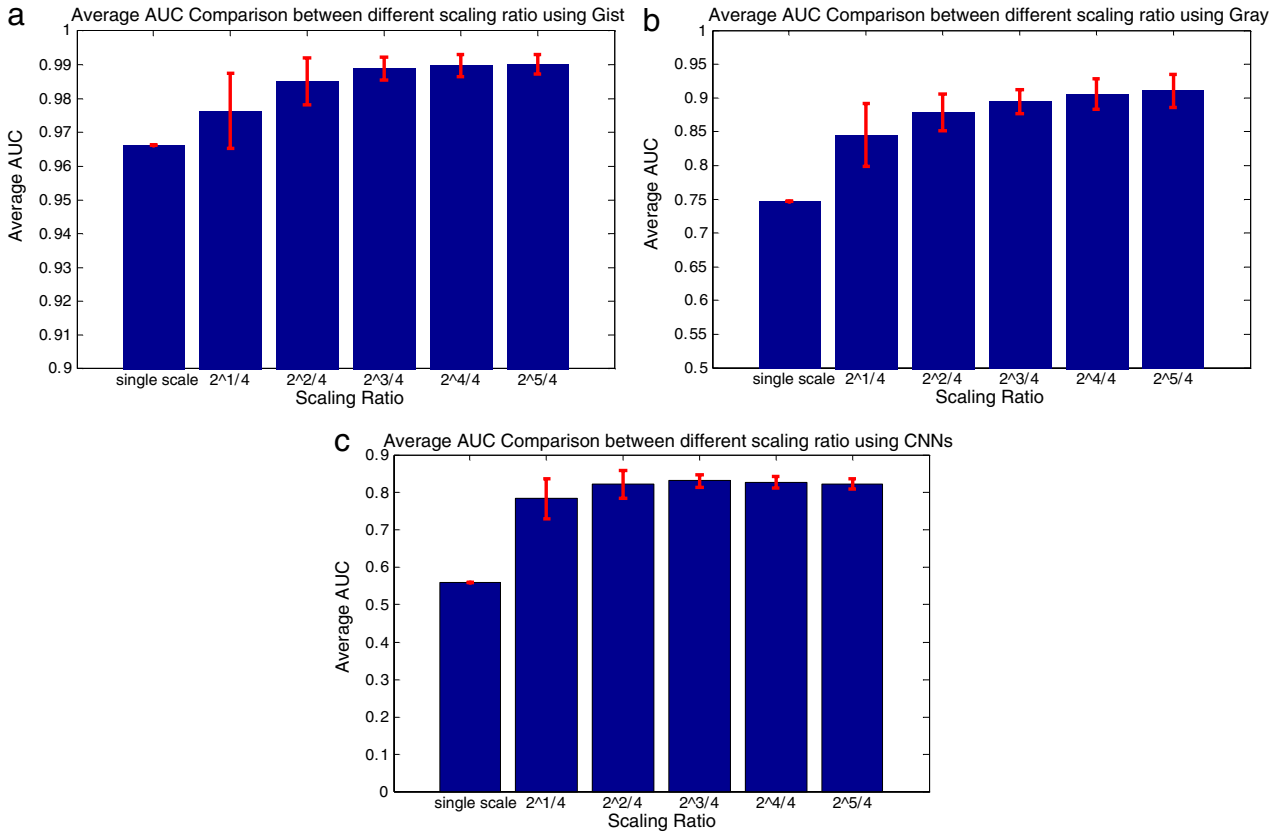

Fig. 13. Average AUC with errorbars for different scaling ratios on the Eynsham dataset using Gist features (a), Gray features (b) and deep learning features (c). $2^{1/4}$ in the x axis indicates the average performance combining scales in a step of $2^{1/4}$.

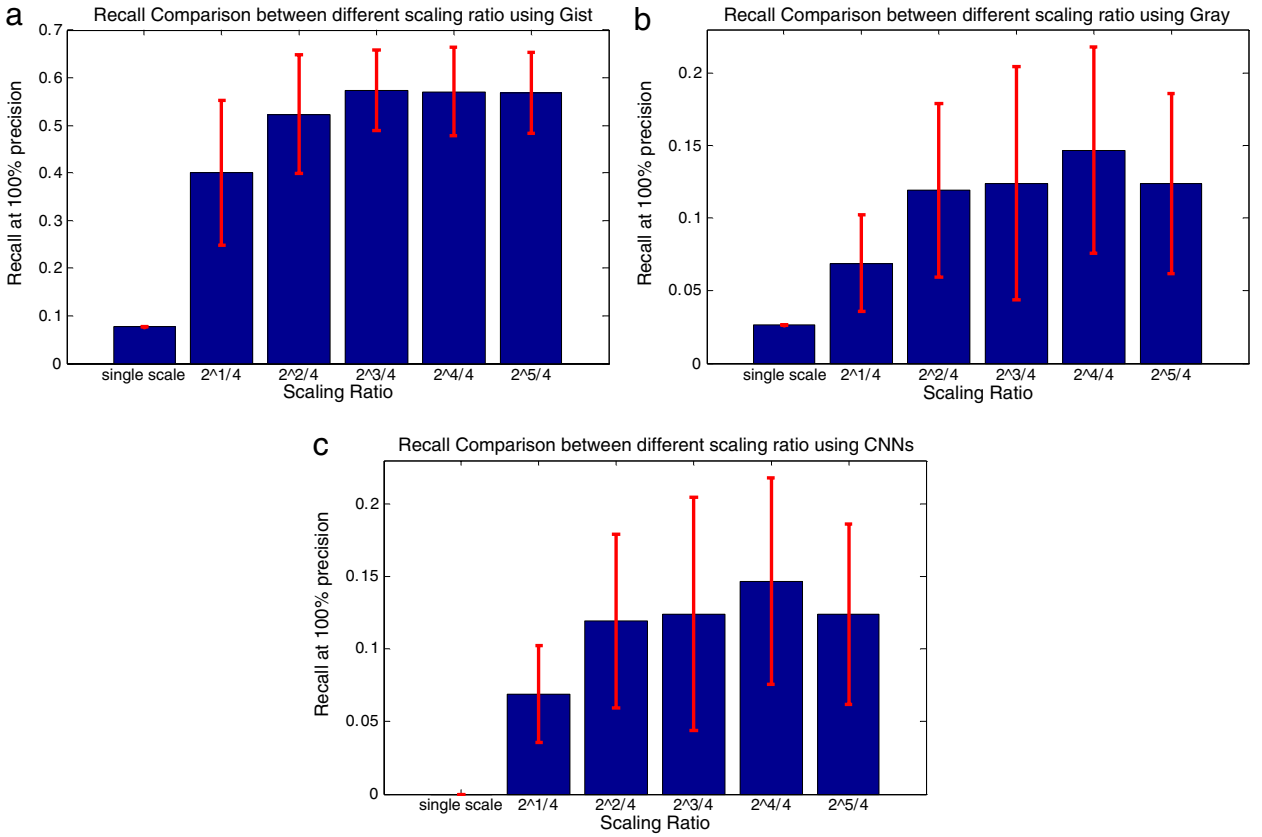

Fig. 14. Maximal recall at 100% precision with errorbars for different scaling ratios on the Eynsham dataset using Gist features (a), Gray features (b) and deep learning features (c). $2^{1/4}$ in the x axis indicates the average performance combining scales in a step of $2^{1/4}$.

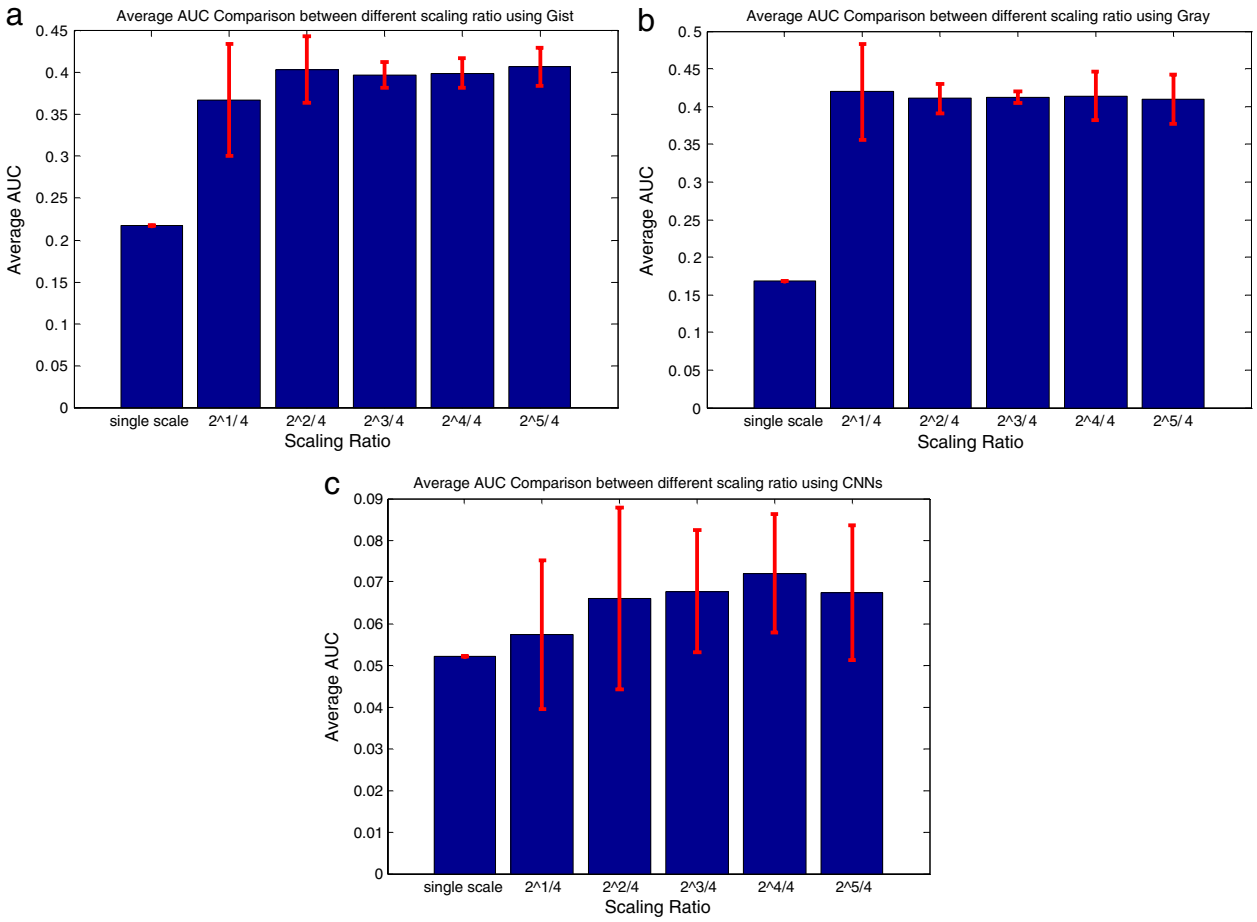

Fig. 15. Average AUC with errorbars for different scaling ratios on the St. Lucia dataset using Gist features (a), Gray features (b) and deep learning features (c). $\mathbf{2^{1/4}}$ in the *x* axis indicates the average performance combining scales in a step of $\mathbf{2^{1/4}}$.

the best performance with the Overfeat feature is achieved when using six scales.

Fig. 12 presents the average AUC when different numbers of combined scales are used on the St. Lucia dataset. On the Gist and Gray features, different numbers of combined scales tend to deliver relatively similar performance, whilst on the Overfeat feature, combining five different scales achieves the best performance which is statistically better than using other number of combined scales.

Figs. 13 and 14 present the effect of different scaling ratios on the AUC and recall at 100% precision on the Eynsham dataset. A similar trend is observed: including a larger coarser scale achieves better performance than using a single scale, for both AUC and recall performance. Scaling ratios of $2^{3/4}$, $2^{4/4}$ and $2^{5/4}$ generate very similar AUC performance, slightly outperforming those using ratios of $2^{2/4}$ and $2^{1/4}$. In Fig. 14, combining multiple scales always delivers statistically better performance than using a single scale, however, the difference in performance improvement between using different number of combined scales is not statistically different.

Fig. 15 describes the influence of different scaling ratios on the St. Lucia dataset, evaluated by the average AUC. The improvement between using different scaling ratios are relatively similar on the Gist and Gray features. On the Overfeat feature, the performance improvement between using multiple scales and using a single scale is not statistically different at 95% confidence level.

5.3. Illustrative multi-scale place recognition combination plots

Fig. 16 illustrates how place match hypotheses at varying scales are combined by a coarser-to-finer localization mechanism and how an incorrect match reported by a single-scale system is corrected by a two-step coarser-to-finer procedure. In general, a large number of false positives at the smallest spatial scale are eliminated due to lack of support from larger spatial scales. The example in Fig. 16 shows how secondary ranked spatially specific matches are correctly chosen as the overall place match due to support from the coarser spatial scales.

6. Conclusions and future work

We have demonstrated that implementing a multi-scale place recognition system improves place recognition performance by combining the output from parallel mapping frameworks, each trained to recognize places at a specific spatial scale. Although this paper presents a suite of specific visual pre-processing techniques and learning mechanisms, we believe that the novel multi-scale combination concept should generalize to other sensor types, sensor processing schemes and learning methods. In this section, we discuss interesting insights from the research and outline several areas of current and future work.

First and foremost, it is clear that incorporating multiple differently-scaled place recognition frameworks in parallel leads to universal performance improvement, regardless of the particular visual processing mechanism or the scaling ratio used. More subtly, there appears to be diminishing returns for having infinite map scales—the performance increase slows rapidly above 3 parallel scales in all experiments and peaks between 4 and 6 scales. Neuroscience experiments have not yet revealed whether there is an upper limit for the number of scales—the results obtained in this

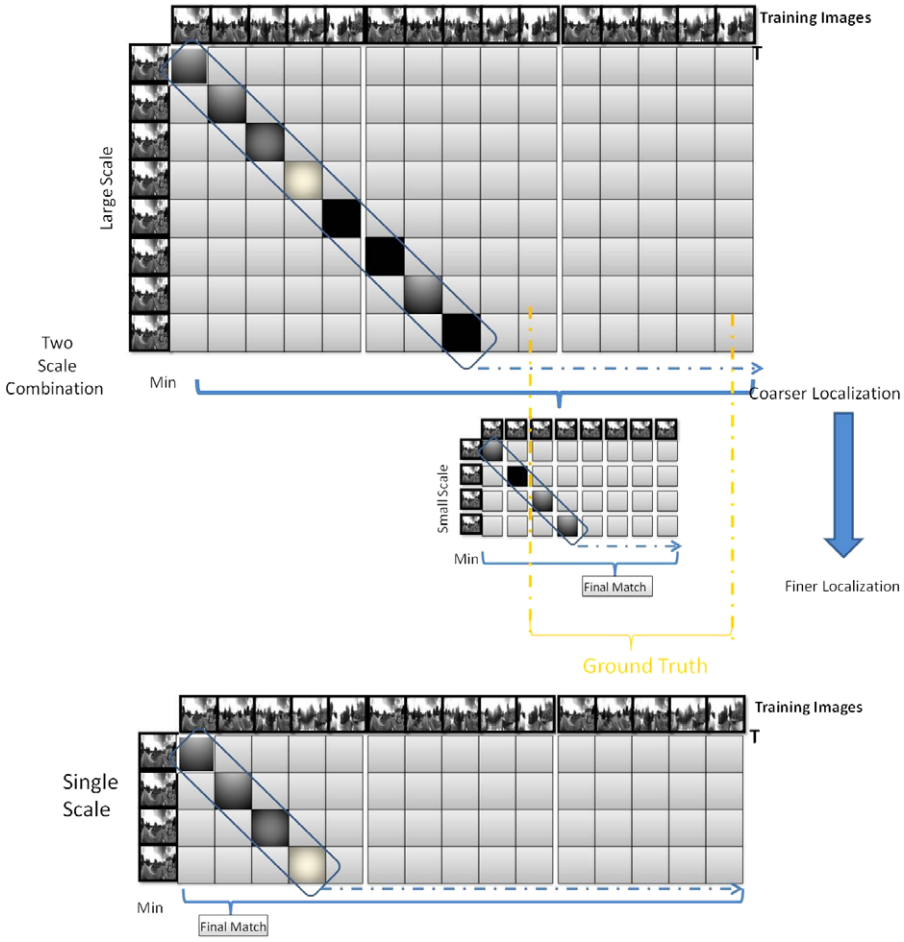

Fig. 16. A two-step coarser-to-finer localization (top) filters out a false positive reported when using a single scale (bottom). The coarser scale generally provides more reliable localization estimates, at which point a finer scale search is performed.

research suggest from a navigation performance basis there is little to gain by having more than a handful of scales. The best scaling ratio, regardless of visual processing mechanism, was a factor of 2 between consecutive scales. As experiments with rodents provide further data, it will become clear whether there is a consistent scaling ratio between consecutive map scales and what the value of this ratio is, although preliminary evidence suggests a scaling ratio of the square root of two.

The current system assumes that the camera is moving at a constant speed during the training and testing stages. To more explicitly introduce odometry as the primary driving source, we incorporated odometry into the segmentation process in a second set of experiments using the St Lucia dataset. In the future, we can extend the use of self-motion to enable the model to expand to two-dimensional unconstrained movement in large open environments. Testing the system in open field environments will be more analogous to many current rodent experiments and may increase the likelihood of generating neuroscience insights.

From a pragmatic perspective, the next step beyond odometry-driven segmentation is data-driven segmentation, where an environment is segmented based on local self-similarity. Such an approach would avoid inefficient representations of large bland spaces with small spatial scale maps. Furthermore, in large open spaces, precise localization is often not possible; in such a situation it may be possible to fall back to a less spatially specific place recognition estimate that uses broader visual cues. To the authors' best knowledge, there is no information on the recruitment density of grid or place cells at different scales in open environments versus cluttered environments—such an experiment would aid further development in this area.

This research has focused on the characteristics of multiple scale maps and their possible benefit for place recognition performance. The mechanism in our model – supervised learning of place segments – is unlikely to be what actually occurs in the mammalian brain. In future work, we will look at implementing the multi-scale mapping framework in a more biologically plausible manner. For example, it may be possible to iterate multiple continuous attractor networks in parallel, each representing a different map scale, and devise a neutrally plausible mechanism by which the ensemble firing of each network can be combined to provide a unique place match hypothesis.

The performance improvement in the result section varies based on the feature extraction methods. This is reasonable because different features capture different semantic information of the images and therefore deliver varying discriminative power. We need to emphasis that the focus of this work is not to propose a learning algorithm that delivers similar performance improvement which is robust to the feature extraction methods used. On the contrary, we focus on the benefit of multiple scale maps and have demonstrated that incorporating multi-scale place recognition hypotheses leads to universal performance improvement, regardless of the feature methods utilized.

Recent work using RatSLAM has shown that biologically inspired algorithms can perform online sensor fusion to enable place recognition in changing environmental conditions, such as over day–night cycles (Jacobson & Milford, 2012; Milford & Jacobson, 2013). An obvious extension to this research would be to use

a multi-scale mapping framework to exploit the variable spatial specificity of different sensor modalities, such as cameras, range finders and WiFi. For example, WiFi signal readings typically provides a much coarser localization signal than a camera and would consequently likely perform better with the coarser network scales (Berkvens, Jacobson, Milford, Peremans, & Weyn, 2014). By integrating these multi-sensor fusion systems with a biologically-inspired, multi-scale mapping framework, it may be possible to combine their functional capabilities to produce a highly capable, general purpose robot mapping and navigation system, finally matching the amazing navigation capabilities of the humble rodent.

Acknowledgments

This work was supported by funding from the Australian Research Council Centre of Excellence CE140100016 in Robotic Vision, a Microsoft Research Faculty Fellowship awarded to MM and an Office of Naval Research ONR MURI N00014-10-1-0936 and Silvio O. Conte Center Grant P50 NIMH MH094263 to MEH.

References

Angeli, A., Filliat, D., Doncieux, S., & Meyer, J.-A. (2008). Fast and incremental method for loop-closure detection using bags of visual words. *IEEE Transactions on Robotics*, *24*(5), 1027–1037.

Bazeille, S., & David, F. (2011). Incremental topo-metric SLAM using vision and robot odometry. In *IEEE international conference on robotics and automation* (pp. 4067–4073).

Berkvens, R., Jacobson, A., Milford, M., Peremans, H., & Weyn, M. (2014). Biologically inspired SLAM using Wi-Fi. In *2014 IEEE/RSJ international conference on intelligent robots and systems*, (IROS 2014). IEEE.

Biederman, I. (1988). Aspects and extension of a theory of human image understanding. In *Computational processes in human vision: an interdisciplinary perspective*.

Bosse, M., Newman, P., Leonard, J., Soika, M., Feiten, W., & Teller, S. (2003). An atlas framework for scalable mapping. In *International conference on robotics and automation*. Taipei, Taiwan: IEEE.

Burak, Y., & Fiete, I. R. (2009). Accurate path integration in continuous attractor network models of grid cells. *PLoS Computational Biology*, *5*(2).

Chen, Z., Jacobson, A., Erdem, U., Hasselmo, M. E., & Milford, M. (2013). Towards bio-inspired place recognition over multiple spatial scales. In *Proceedings of the 2013 Australasian conference on robotics & automation* (pp. 1–9).

Chen, Z., Jacobson, A., Erdem, U. M., Hasselmo, M. E., & Milford, M. (2014). Multi-scale bio-inspired place recognition. In *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE.

Chen, Z., Obadiah, L., Jacobson, A., & Milford, M. (2014). Convolutional neural network based place recognition. In *Australiaian conference on robotics and automation. Melbourne, Australia*.

Cummins, M., & Newman, P. (2008). FAB-MAP: probabilistic localization and mapping in the space of appearance. *International Journal of Robotics Research*, *27*(6), 647–665.

Cummins, M., & Newman, P. (2009). Highly scalable appearance-only SLAM—FAB-MAP 2.0., In *Robotics: science and systems, Seattle, United States*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition, 2009*, CVPR 2009. IEEE.

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, *463*(7281), 657–661.

Erdem, U. M., & Hasselmo, M. E. (2014). A biologically inspired hierarchical goal directed navigation model. *Journal of Physiology-Paris*, *108*(1), 28–37.

Erdem, U. M., Milford, M. J., & Hasselmo, M. E. (2015). A hierarchical model of goal directed navigation selects trajectories in a visual environment. *Neurobiology of Learning and Memory*, *117*, 109–121.

Giovannangeli, C., Gaussier, P., & Désilles, G. (2006). Robust mapless outdoor vision-based navigation. In *2006 IEEE/RSJ international conference on intelligent robots and systems*. IEEE.

Girdhar, Y., & Dudek, G. (2012). Efficient on-line data summarization using extremum summaries. In *IEEE international conference on robotics and automation, ICRA*.

Glover, A. J., Maddern, W. P., Milford, M. J., & Wyeth, G. F. (2010). FAB-MAP + RatSLAM: Appearance-based SLAM for multiple times of day. In *International conference on robotics and automation*. Anchorage, United States: IEEE.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005a). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *11*(436), 801–806.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005b). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806.

Jacobs, J., Weidemann, C. T., Miller, J. F., Solway, A., Burke, J. F., Wei, X.-X., & Fried, I. (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature Neuroscience*, *16*(9), 1188–1190.

Jacobson, A., & Milford, M. (2012). Towards brain-based sensor fusion for navigating robots. In *Proceedings of the 2012 Australasian conference on robotics & automation*. Australian Robotics & Automation Association.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer.

Killian, N. J., Jutras, M. J., & Buffalo, E. A. (2012). A map of visual space in the primate entorhinal cortex. *Nature*, *491*(7426), 761–764.

Konolige, K., Marder-Eppstein, E., & Marthi, B. (2011). Navigation in hybrid metric-topological maps. In *2011 IEEE international conference on robotics and automation (ICRA)*. IEEE.

Kuipers, B. (1978). Modeling spatial knowledge. *Cognitive Science*, *2*(2), 129–153.

Kuipers, B. (2000). The spatial semantic hierarchy. *Artificial Intelligence*, *119*(1), 191–233.

Kuipers, B., & Byun, Y. T. (1991). A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Robotics and Autonomous Systems*, *8*(1), 47–63.

Kuipers, B., Modayil, J., Beeson, P., MacMahon, M., & Savelli, F. (2004). Local metrical and global topological maps in the hybrid spatial semantic hierarchy. In *International conference on robotics and automation, New Orleans, USA*.

Milford, M. (2013). Vision-based place recognition: How low can you go? *International Journal of Robotics Research*, *32*(7), 766–789.

Milford, M., & Jacobson, A. (2013). Brain-based sensor fusion for navigating robots. In *IEEE international conference on robotics and automation*. Karlsruhe, Germany: IEEE.

Milford, M., & Wyeth, G. (2008). Mapping a suburb with a single camera using a biologically inspired SLAM system. *IEEE Transactions on Robotics*, *24*(5), 1038–1053.

Milford, M., & Wyeth, G. (2010). Persistent navigation and mapping using a biologically inspired SLAM system. *International Journal of Robotics Research*, *29*(9), 1131–1153.

Milford, M., & Wyeth, G. (2012). SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *IEEE international conference on robotics and automation*. St Paul, United States: IEEE.

Milford, M. J., Wyeth, G., & Prasser, D. (2004). RatSLAM: A hippocampal model for simultaneous localization and mapping. In *IEEE international conference on robotics and automation*. New Orleans, USA: IEEE.

Newman, P., Cole, D., & Ho, K. (2006). Outdoor SLAM using visual appearance and laser ranging. In *International conference on robotics and automation*. Florida, United States: IEEE.

O'Keefe, J., & Conway, D. H. (1978). Hippocampal place units in the freely moving rat: Why they fire where they fire. *Experimental Brain Research*, *31*(4), 573–590.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely moving rat. *Brain Research*, *34*(1), 171–175.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Paul, R., Rus, D., & Newman, P. (2012). How was your day? Online visual workspace summaries using incremental clustering in topic space. In *2012 IEEE international conference on robotics and automation, ICRA*.

Potter, M. C. (1975). Meaning in visual search. *Science*, *187*, 965–966.

Rowekamper, J., Sprunk, C., Tipaldi, G. D., Stachniss, C., Pfaff, P., & Burgard, W. (2012). On the position accuracy of mobile robot localization based on particle filters combined with scan matching. In *2012 IEEE/RSJ international conference on intelligent robots and systems, IROS*.

Schindler, G., Brown, M., & Szeliski, R. (2007). City-scale location recognition. In *IEEE conference on computer vision and pattern recognition, 2007. CVPR'07*.

Segvic, S., Remazeilles, A., Diosi, A., & Chaumette, F. (2009). A mapping and localization framework for scalable appearance-based navigation. *Computer Vision and Image Understanding*, *113*(2), 172–187.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. ArXiv Preprint arXiv:1312.6229.

Stensola, H., Stensola, T., Solstad, T., Froland, K., Moser, M., & Moser, E. (2012). The entorhinal grid map is discretized. *Nature*, *492*(7427), 72–78.

Taube, J. S., Muller, R. U., & Ranck, J. B. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *The Journal of Neuroscience*, *10*(2), 420–435.

Theocharous, G., Murphy, K., & Kaelbling, L. P. (2004). Representing hierarchical POMDPs as DBNs for multi-scale robot localization. In *2004 IEEE international conference on robotics and automation, 2004. Proceedings*, ICRA'04. IEEE.

Ulanovsky, N., & Moss, C. F. (2007). Hippocampal cellular and network activity in freely moving echolocating bats. *Nature Neuroscience*, *10*(2), 224–233.

Ulrich, I., & Nourbakhsh, I. (2000). Appearance-based place recognition for topological localization. In *IEEE international conference on robotics and automation, San Francisco, CA, USA*.

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2005). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*.

Welinder, P. E., Burak, Y., & Fiete, I. R. (2008). Grid cells: the position code, neural network models of activity, and the problem of learning. *Hippocampus*, *18*(12), 1283–1300.

Yartsev, M. M., Witter, M. P., & Ulanovsky, N. (2011). Grid cells without theta oscillations in the entorhinal cortex of bats. *Nature*, *479*(7371), 103–107.