Bio-inspired multi-scale fusion

Stephen Hausler · Zetao Chen · Michael E. Hasselmo · Michael Milford

Received: 27 November 2019 / Accepted: 27 March 2020 / Published online: 22 April 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
We reveal how implementing the homogeneous, multi-scale mapping frameworks observed in the mammalian brain’s mapping systems radically improves the performance of a range of current robotic localization techniques. Roboticists have developed a range of predominantly single- or dual-scale heterogeneous mapping approaches (typically locally metric and globally topological) that starkly contrast with neural encoding of space in mammalian brains: a multi-scale map underpinned by spatially responsive cells like the grid cells found in the rodent entorhinal cortex. Yet the full benefits of a homogeneous multi-scale mapping framework remain unknown in both robotics and biology: in robotics because of the focus on single- or two-scale systems and limits in the scalability and open-field nature of current test environments and benchmark datasets; in biology because of technical limitations when recording from rodents during movement over large areas. New global spatial databases with visual information varying over several orders of magnitude in scale enable us to investigate this question for the first time in real-world environments. In particular, we investigate and answer the following questions: why have multi-scale representations, how many scales should there be, what should the size ratio between consecutive scales be and how does the absolute scale size affect performance? We answer these questions by developing and evaluating a homogeneous, multi-scale mapping framework mimicking aspects of the rodent multi-scale map, but using current robotic place recognition techniques at each scale. Results in large-scale real-world environments demonstrate multi-faceted and significant benefits for mapping and localization performance and identify the key factors that determine performance.

Keywords Place recognition · Biologically inspired navigation · Aerial navigation

1 Introduction
Decades of research into the neural mechanisms underlying mapping have revealed a range of neurons with strong spatial correlates in various regions of the mammalian brain. Much of the early research focused on the brain region known as the hippocampus in rodents, revealing a type of cell called a “place cell” (O’Keefe and Conway 1978). Subsequent research revealed a range of additional navigation-related cells including head-direction cells (Taube et al. 1990) (which respond based on the orientation of the animal’s head), boundary cells (Lever et al. 2009) (which respond to physical barriers in the environment) and grid cells (Hafting et al. 2005). Each grid cell encodes not one, but a number of locations in the environment, arranged in a hexagonal tessellating pattern, or lattice, over the environment. The spacing between firing fields of these grid cells varies from the dorsal to the ventral parts of the entorhinal cortex.
ventral region, creating multiple grid cell patterns with different spatial scales (Sargolini et al. 2006).

The homogeneous, discrete multi-scale map representation exists in stark contrast to typical robotic mapping systems, which usually encode the world at one, or at most two spatial scales. Furthermore, when using two map scales, robotic mapping approaches typically combine a local “metric” map of space with a global topological map of the connectivity of space (Bosse et al. 2003; Kuipers 2007). The deliberate aliasing of associations between a single grid cell and potentially thousands of locations in the environment is also very different from robotic approaches, where the focus has been on solving the data association problem to avoid this very issue.

In this research, we investigate the combination of homogeneous multi-scale mapping and localization systems and a range of state-of-the-art approaches to place representation, in two substantially different large-scale environments. Our first environment is an aerial dataset, navigating using downward facing images, while our second environment is a forward-facing navigation experiment on a train route. Rather than considering each discrete motion of an autonomous robot as a single place, we learn place representations at each location across nine different spatial scales, spanning a total spatial scale range of up to 16 times the smallest scale. Larger scales consider visual information from a large number of temporally adjacent small-scale locations. Place representations at various scales are matched to all equivalent place representations throughout the entire map database, before being combined to produce a global location hypothesis. We also mimic the periodicity found within rodent grid cells, by finding multiple potential location hypotheses at the smaller scales within small spatial regions. These periodic potential best matching location hypotheses are filtered using larger scales with larger spatial regions. A high-level overview diagram showing our approach concept, for our aerial dataset, can be seen in Fig. 1. A similar method diagram is provided for our ground dataset in Fig. 2. Our approach to multi-scale navigation is uniquely different to prior work in this space; in a previous approach, scales were created by clustering different sized sets of images using an LMNN (Chen et al. 2015). Any clustering algorithm is designed to find a representation approximating the total visual information within the cluster; however, a more powerful representation is to directly use all the visual information within that spatial scale (with the downside of increased computational cost). Our inspiration for combining multiple scales stems from prior work combining different image processing methods together (Hausler et al. 2019). We expand the concept of fusing information from different image representations by method to instead fusing information from multiple spatial sizes within the observed visual scene.

We create place representations using a selection of different techniques within the computer vision and robotics communities. We utilize a learnt feature method (NetVLAD) (Arandjelovic et al. 2018), a bio-inspired hand-crafted feature (Gist) (Oliva and Torralba 2006) and direct pixel-based encodings (Sum of Absolute Differences) (Milford and Wyeth 2012), with a different set of experiments for each type of place representation.

We evaluate our theorems for both a simulated aerial navigation traverse and a ground vehicular traverse. Our simulated aerial traverse uses downward facing aerial images extracted from Nearmap (Nearmaps 2019; Mount et al.
2 Background

Rodents possess a remarkable ability to navigate in a diverse range of environments, and significant research has been performed to investigate the neural mechanisms behind the successful navigation strategies of rodents. A key finding was the discovery of place cells in the hippocampus (O’Keefe and Conway 1978). Place cells activate when a rodent is in a pre-learnt location in the environment, although they exhibit a range of experience, environment and behavior-modulated influences as well (Schoenenberger et al. 2016). Later on, a different type of cell called a grid cell was discovered in the medial entorhinal cortex (Hafting et al. 2005). The spatial firing fields associated with a grid cell are structured in a hexagonal lattice, and the size of firing fields and spacing between locations within the hexagonal lattice varies between different sub-populations of grid cells, from centimeters to at least several meters (Brun et al. 2008), with a typical ratio between consecutive map scales of approximately $\sqrt{2}$ (Stensola et al. 2012). The upper limit of grid cell map scales is unknown, in significant part because of the technical difficulty of recording sufficient data from freely roaming rodents in large environments. There is also evidence that the rodent brain simultaneously uses multiple grid cell scales when navigating. It is hypothesized that spatial location can only be decoded from grid cells when multiple spatial scales are combined, since grid cells fire in a periodic fashion with firing fields closer than the dimensions of the environment (Sreenivasan and Fiete 2011; Stemmler et al. 2015).

The existence of cells in the brain with strong spatial correlates and different properties to conventional robotic mapping systems has naturally been a source of significant interest to the robotics community. The combination of place and head-direction cells was used to formulate a range of robotic mapping systems (Arleo and Gerstner 2000; Browning 2000; Krichmar et al. 2005) including RatSLAM (Milford et al. 2004), a biologically inspired simultaneous localization and mapping (SLAM) algorithm which performed path integration and encoded learnt memories using continuous attractor network models of place cells, head-direction cells and grid cells (Milford and Wyeth 2008, 2010; Yu et al. 2019). A simulation of a continuous attractor network demonstrated the natural generation of grid cell characteristics in a network used for path integration (Burak and Fiete 2009), while further simulations have spontaneously generated grid cells (Banino et al. 2018). Furthering this concept, the grid cell model has been used to design both a prototype multi-scale navigation model (Erdem and Hasselmo 2014) and a multi-scale localization system (Chen et al. 2015), exploiting the hierarchical and hexagonal structure of neural responses in the medial entorhinal cortex (MEC). The work of Chen et al. clusters image features across different spatial scales, whereas in this work we directly increase the visual information provided to larger spatial scales. The aliasing of one grid cell to many locations has also been explored as a means of forming highly efficient map representations, both in theory (Burak and Fiete 2009; Sreenivasan and Fiete 2011) and in practice in robotics (Yu et al. 2017; Jacobson et al. 2017).

What all existing neuroscience and robotic studies lack has been the ability to deploy modern, state-of-the-art approaches to place recognition in an environment where the true benefits of homogeneous mapping over a large range of scales can be evaluated. In biology, we know from behavioral observations that wild animals can roam over very large areas (Moser 2011), but our ability to record spatial neuronal activity is limited by both technical considerations and an inability to sufficiently sample over such long paths (compared to repeated visits to locations in a small 1 by 1 meter arena, as is typical in laboratory experiments). In robotics, research has primarily focused on high-accuracy 3D metric mapping and localization in environments of limited scale or topological complexity (Mur-Artal et al. 2015), or on mapping and navigation in larger environments that are predominantly path-like in nature (Furgale and Barfoot 2010), such as in autonomous vehicle applications (Cummins and Newman 2009). Beyond the limited one- or two-scale approaches already noted, the majority of current techniques also process the world in either a frame-by-frame manner, or as a global topological or metric map: the concept of place representations of all different scales being absent from most approaches. Only recently has global imagery over a vast range of both time (many years and times of day) and spatial scales (from centimeters to thousands of kilometers) become available, enabling us to investigate the utility of the...
For our ground-based dataset and method, we aggregate visual information using temporal sequences of images. On the left, we have the reference dataset in summer, which is copied and aggregated into sets of temporal sequences of varying lengths. In other words, for each image in summer, multiple scales are created by aggregating image features looking backwards in time, with different sequence aggregation lengths depending on the scale size. Each aggregated set is normalized to an equal feature vector length (400 dimensions). The right side shows the current query in winter and a collection of frames looking earlier in time. Sequences of equal length are compared, with the query sequence compared against all reference sequences and for all scale sizes. (This diagram shows four scales, but the actual algorithm can have up to nine scales.)

multi-scale homogenous mapping and localization systems observed in the brain in challenging, real-world environments.

3 Approach

In place recognition, the objective is to successfully match the currently viewed scene received by a camera system to a stored collection of memories, typically referred to as a “reference database” or “map.” This is often termed the “data association” problem, and successfully completing this task is required for loop-closure in autonomous navigation. To compare the current scene to the reference images, typically some manner of feature extraction and dimensionality reduction is used, to improve both the matching speed while discarding distracting image features. In this work, we use the well-established image processing methods of Sum of Absolute Differences with Patch Normalization (Milford and Wyeth 2012), Gist (Oliva and Torralba 2006), and NetVLAD (Arandjelovic et al. 2018), which collectively span the range of technique types used in the field of robotics (Lowry et al. 2016): a direct image-based approach (SAD), a handcrafted feature detector (Gist) and a learnt feature detector (NetVLAD). Every reference image and each new scene are then converted to an encoded representation of the original image, and image comparison is performed using the Euclidean distance metric.

3.1 Algorithm for aerial navigation

We provide two slightly different algorithms in this paper, one for the aerial dataset and the other for the ground dataset. Different algorithms are required since aerial images have different properties compared to forward-facing images (such as planar versus non-planar). For our aerial algorithm, we begin by creating a collection of different scaled images from an original, wide field-of-view birds-eye image. Each scale is a center-cropped view of the previous larger scale, and the largest scale is the original image (see Fig. 3). All scales are then resized to the input size of the image processing method, which is 224 by 224 pixels for NetVLAD, 128 by 128 pixels for Gist and 32 by 64 pixels for SAD. This crop-then-resize effect creates multiple collections of images—the largest scale observes the full field-of-view of the scene, but at a low resolution. The smallest scale observes a small field-of-view of the scene, but at a high resolution.

The creation of multiple scales is performed for both the reference database images and the query images. We then compare the current query location to all the database locations, separately comparing individual scales (where each scale is a different image of the same location). For example, the smallest scale image in the query location is compared against all of the smallest scale images in the database. This comparison produces a difference vector for each scale, where the values within the difference vector are the Euclidean distance from the current query scale image to

 Springer
Fig. 3  Montage of the nine aerial scales, from smallest in the top left to largest in the bottom right, showing rural scenery (a) and urban scenery (b). The multi-scale implementations used here use combinations of these snapshots. For example, a two-scale approach uses the top left plus the bottom right place representations, while a nine-scale approach uses all nine place representations, all centered on the smallest scale view.

At this stage, we could simply sum the difference vectors across $s$; however, this would be an unbalanced operation as the distribution of difference scores varies between scales. Inspired by both SeqSLAM (Milford and Wyeth 2012) and the periodicity of grid cell firing fields, we find periodic local best matching images within the difference scores of each scale. Normally, the minimum score in $D_s$ would be the globally best matching image (for the current query $q_s$) out of all reference images, for that scale size. Instead, we find multiple competing local best matches across $l$ local regions where each region consists of $K$ reference images. $K$ can be interpreted as the spatial period of our search, and we vary $K$ by scale size, where the smallest scales are highly periodic and the largest scale is not periodic at all. Or in other words, we search for the global best matching location at the largest scale. We create these periodic match candidates by applying local normalization to each region $l$, where each region is centered so that the difference scores in that region have a mean of zero and standard deviation of one.

$$D_s = \sqrt{\sum_{i=1}^{N} (q_s(i) - r_s(i))^2} \quad \forall s$$

where $q_s$ is the current query image representation (from either NetVLAD, Gist or SAD) from scale $s$, $r_s$ is the set of reference image representations from scale $s$, and $N$ is the number of database images.

At this stage, we could simply sum the difference vectors across $s$; however, this would be an unbalanced operation as the distribution of difference scores varies between scales. Inspired by both SeqSLAM (Milford and Wyeth 2012) and the periodicity of grid cell firing fields, we find periodic local best matching images within the difference scores of each scale. Normally, the minimum score in $D_s$ would be the globally best matching image (for the current query $q_s$) out of all reference images, for that scale size. Instead, we find multiple competing local best matches across $l$ local regions where each region consists of $K$ reference images. $K$ can be interpreted as the spatial period of our search, and we vary $K$ by scale size, where the smallest scales are highly periodic and the largest scale is not periodic at all. Or in other words, we search for the global best matching location at the largest scale. We create these periodic match candidates by applying local normalization to each region $l$, where each region is centered so that the difference scores in that region have a mean of zero and standard deviation of one.

$$D_{z,l}(i) = \frac{D_s(i) - \mu_s(l)}{\sigma_s(l)} \quad \forall i \quad \forall s$$

where $\mu_s(l)$ is the local mean in region $l$, $\sigma_s(l)$ is the local standard deviation in region $l$, and the period is $K$. Our values of $K$ are 10, 25, 50, 100, 200, 400, 800, 1600 and $N$ for each successive scale size. Equation 2 produces a normalized difference score for each scale, with a multitude of local best matches at the smaller scales. The overall best candidate is the most agreed-upon candidate location over these multiple scales. We combine these difference scores at different scales by summing the difference score at each scale, as expressed in Eq. 4:

$$D_{all} = \sum_{n=1}^{S} D_{z,n}$$

where $S$ is the number of scales. After producing the combined difference vector, we perform a second normalization procedure, to center the data on a mean of zero and standard deviation of one, as shown in Eq. 5:

$$D_z = \frac{D_{all} - \bar{D}_{all}}{\sigma(D_{all})}$$

The best place match is then calculated by finding the minimum difference score in the normalized combined difference vector $D_z$:

$$\text{match}(i) = \arg\min_{i \in N} (D_z)$$
The difference score at this minimum is used as a quality score—the further this score is from 0 in the negative direction, the better the match estimate. During deployment, the quality score is used to decide whether the agent is in a new, unknown location, or somewhere it has been before:

$$\text{localize}(i) = \begin{cases} 1, & \min(D_z) < \text{thresh} \\ 0, & \min(D_z) \geq \text{thresh} \end{cases} \quad (7)$$

As per standard evaluation processes, we vary the threshold for this decision to generate a precision–recall curve. Precision is the proportion of localization events that are correct, while recall is the proportion of times the agent decides it is in a place it has been before. Precision and recall can be calculated from the number of true positives (correct localizations), false positives (incorrect localizations), and false negatives (missed localizations), as expressed in the following equations:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

By varying our localization decision threshold, we produce a precision–recall curve from which we calculate the area under this curve (AUC), which is a measure of the overall localization capability of the approach:

$$\text{AUC} = \int_{\text{Recall}=0}^{\text{Recall}=1} \text{Precision(Recall)} \quad (10)$$

We also provide our source code on Github, available at: https://github.com/StephenHausler/Multi-Scale-Fusion.

### 3.2 Algorithm for ground navigation

For the aerial dataset, larger scales are a view of a location from a higher altitude, thus observing a larger amount of visual information around the current location of the autonomous robot. The same principle can be approximated for a ground, forward-facing vehicle by concatenating image features over a temporal sequence of images. While ground and aerial images are quite different in their visual structure (such as planar versus non-planar), the concept of providing visual information that corresponds to varying spatial scale coverage (whether from a larger field-of-view or a longer sequence of images) is equally applicable. The ground dataset does have reduced efficiency in this process due to visual overlap in forward-facing imagery; however, we propose that the same algorithms can be used for both methods with only the image pre-processing varying, as is the case in nature with different sensing modalities and pre-processing streams.

Different scales are produced by using different length concatenation sequences, which aggregates feature vectors from both reference and query traverses into longer feature vectors. This process creates 6 copies of the original dataset, where each element of the database is converted into an aggregated feature vector using past frames (see Fig. 4 for more detail). To prevent excessive computation requirements with this approach, we apply significant dimensionality reduction to reduce each feature vector to a standardized length of 400 frames. We use a computationally cheap dimensionality reduction approach, by downsampling the concatenated feature vectors. Specifically, every $i$th element of the concatenated feature vector is used to form a shorter feature vector with a dimension of 400 (we use the same dimension for all three image processing methods and for all scale sizes). Our proposed approach of creating sequential clusters differs from previous work by the authors, since earlier work used the mean sensory representation (Jacobson et al. 2018) over a place cell instead. We also expand from the prior work using place cells to a grid cell representation, with the inclusion of multiple spatial scales.

Once the collection of ground sequenced features is created, the same equations are used as in the aerial algorithm. To recap, each scale consists of a 400-dimension feature vector, which is a down-sampled version of a much longer feature vector which contains visual information over a temporal sequence of images. We again use the Euclidean distance metric to match each aggregated query feature vector to each and every aggregated reference feature, for every scale (see Eq. 1). Normalization is performed for each scale separately, using the same periodic normalization process. Please refer to Eqs. 2 and 3 from Sect. 3.1. As our ground dataset is much larger than our aerial dataset, we use larger values of $K$ for successive spatial periods: 30, 60, 120, 250, 500, 1000, 2000, 4000 and $N$. We then summate these normalized difference scores across the number of scales (see Eq. 4). Finally, we apply Eqs. 5–10 to evaluate the localization performance of the ground dataset.

### 3.3 Scale selection

The choice of scales was inspired by research on grid cells by Stensola and colleagues (Stensola et al. 2012), who found that grid cells in the rodent hippocampus appeared to increase their field size at a factor of approximately $\sqrt{2}$ per “level” of the grid cell lattice. Additionally, for the aerial dataset, this ratio causes each successive scale to encode an aerial area double that of the previous scale (doubling the amount of visual input per scale). Sections 5 and 6 will discuss the optimal scale ratio in more detail.

On the Nearmap aerial dataset, we selected the smallest scale to be a real-world size of 56 by 56 m. Therefore, for nine spatial scales, the sizes of each successive scale are approx-
Fig. 4 In this tabular diagram, we visually demonstrate the procedure for creating our concatenated sequences from currently viewed scenes (for the ground method). As each new image is observed by the autonomous agent, the feature representations are binned into one or more sequence lengths. Longer sequence lengths concatenate the current feature vector with feature vectors from past frames. This diagram displays the raw images for improved comprehension, while the actual algorithm uses feature vectors instead.

| Seq. Length (scale size) | Query 1 | Query 2 | Query 3 | Query 4 | Query 5 | Query 6 |
|-------------------------|---------|---------|---------|---------|---------|---------|
| 1                       | ![Image](image1) | ![Image](image2) | ![Image](image3) | ![Image](image4) | ![Image](image5) | ![Image](image6) |
| 2                       | ![Image](image7) | ![Image](image8) | ![Image](image9) | ![Image](image10) | ![Image](image11) | ![Image](image12) |
| 3                       | ![Image](image13) | ![Image](image14) | ![Image](image15) | ![Image](image16) | ![Image](image17) | ![Image](image18) |
| 4                       | ![Image](image19) | ![Image](image20) | ![Image](image21) | ![Image](image22) | ![Image](image23) | ![Image](image24) |

imately as follows: 79, 112, 158, 224, 316, 448, 633 and 896 m. The size of the smallest scale was an experimental heuristic, based on the number of observable visual features to a human observer. We experimented with different minimum scales in our absolute scale size evaluation experiment in Sect. 5.4, with smallest scales at 56, 59, 63, 67, 71, 75 and 79 m.

On the Nordland ground dataset, we maintain a scale ration of approximately $\sqrt{2}$, with the smallest scale being a single image in the traverse. Successive scales are the sequential concatenation of image features, with sequence sizes as follows: 1, 2, 3, 4, 6, 8, 11, 16 and 23 images. The real-world spacing of images in our dataset is approximately 20 m; thus, these scale sizes in real-world distances are as follows: 0 m, 20 m, 40 m, 60 m, 100 m, 140 m, 200 m, 300 m and 440 m (similar to the aerial dataset).

### 3.4 Statistical analysis

In our results, we perform statistical analysis using student’s $t$ test to observe if there is a statistical difference between different scales when multiple navigation trials are used. The $t$ test score, comparing two sets of AUC values, is calculated using Eq. 11:

$$t = \frac{AUC_2 - AUC_1}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}} \quad \text{(11)}$$

where $s_1^2$ denotes the standard deviation of the set of values in $AUC_1$, while $s_2^2$ denotes the standard deviation of the set of values in $AUC_2$. $n$ is the number of navigation trials, which, for our experiments, is the same across all scale combinations. The calculated $t$ value is then compared against a $t$ value table, for $2n - 2$ degrees of freedom and $p < 0.05$ significance. If the calculated $t$ value is greater than the value in the table, we reject the null hypothesis and instead use the alternative hypothesis.

### 4 Experimental setup

#### 4.1 Theoretical background to experimental methodology

Our experiments and results are aimed at identifying the functional performance effects of combining a collection of different spatial scales, where a spatial scale defines a collection of accumulated visual information with respect to the position of the robot or sensor. This accumulation comes from a larger field-of-view for the aerial dataset or the observation of many sequential images in the case of the ground dataset. A larger region will encompass visual features from multiple smaller regions (distal cues), while the smallest region only includes the visual features that are most relevant to the current position (proximal cues). Intuitively, the smaller
region would have the highest positional accuracy; however, a low number of visual features increases the risk of perceptual aliasing in large datasets. This is because there are less permutations in visual features available, particularly if the region-of-view consists of large, bland visual features. In terms of rodent navigation, recent experimental research has suggested that rodents can use either distal and proximal cues, with a preference toward distal cues (Hébert et al. 2017). A preference towards distal cues, which is similar to using a larger spatial scale, suggests that perceptual aliasing is a problem which is naturally solved within the rodent brain.

### 4.2 Aerial dataset

Our aerial dataset is a concatenation of images recorded from different regions within Australia at two different times, one set in 2013, the other in 2019, extracted from Nearmap satellite imagery (Nearmaps 2019). In this 6-year period, the visual appearance and structure of the scenery has changed significantly. This is due to both time of day factors and environmental changes, such as urban construction or the crop cycle of a farm. Together this dataset has for the first time the key properties critical for investigating the utility of a homogeneous, multi-scale mapping and localization approach in a way that existing animal experimentation and existing robotic datasets and simulation cannot.

Nearmap imagery for a given place representation was downloaded using the Nearmap API, capturing an overview image slightly larger than the largest chosen scale for that place. The images are also rotated to be oriented in a specific pre-defined direction of travel within the world-coordinate frame. This ensures that the images are very similar to the images that would be captured by a high-altitude UAV with the directional information available from a sensor such as a compass. Successive downloaded images are separated by a mixed random and preset distance between them, in terms of GPS distance between image centroids. This distance is always between 10 and 20 m. Adding some variation in the distance between images adds an additional aspect of realism, since most autonomous systems will have some velocity variations across their navigation. We also add a small perpendicular (to the direction of travel) offset between the traverses recorded at different times, again to improve realism. To produce multiple scales, the original Nearmap images are center-cropped by varying amounts, producing \( S \) (number of scales) cropped copies of the original image, which then go through the image processing described at the start of this section. As larger scales view an area larger than the distance between image centroids, reference images at large scales will overlap. Overlap causes location ambiguity in the localization process, requiring smaller scales to find the specific matching location.

### Table 1 Aerial dataset specifications and parameters

| Specification                  | Value        |
|-------------------------------|--------------|
| GPS tolerance for AUC graphs  | 50 m         |
| Total number of reference set images | 3041         |
| Number of query images per trial | 500         |
| Number of navigation trials   | 20           |
| Real-world distance between successive images | 10–20 m      |
| Image resize for SAD          | 32 by 64 pixels |
| Image resize for Gist         | 128 by 128 pixels |
| Image resize for NetVLAD      | 224 by 224 pixels |

In each navigation trial, we randomly select 500 images from the 2019 set of images (hereafter we define this as the “query set”) and compare this set against all images from 2013 (the “reference set”). Our query set is shuffled; thus, each following image can be any distance from the current image. Such an approach, while unlikely to occur in a real traverse, provides the greatest proof of concept that the approach is not reliant on temporal information and is purely observing the visual features over multiple spatial scales. For each navigation trial, we evaluate our multi-scale technique observing the visual features over multiple spatial scales. For the approach in a way that existing animal experimentation and existing robotics datasets and simulation cannot.

### 4.3 Ground dataset

For our ground dataset, we use the publicly available Nordland dataset (Sünderhauf et al. 2013), which is recorded from a train traveling through Norway across different seasons (summer, autumn, winter and spring). The Nordland dataset is considered a benchmark dataset in visual place recognition, with different sections of the full recording used in a wide variety of related literature (Naseer et al. 2018; Hausler et al. 2020; Sünderhauf et al. 2015).

In this work, we use the summer traverse as our reference database of images, and we attempt to localize against this dataset using images recorded during winter. We extract 8280 images at 1 FPS from each condition from the original videos. Our extraction is frame-aligned, such that the first image in summer is the same location as the first image in winter, and so on. At an average spacing of 20 m between frames, this corresponds to a traverse of approximately 165km. For our localization experiments, again we select 500 random locations (images) from the winter set and attempt to localize each image with respect to the reference dataset. For each of our 500 locations, we also extract a sequence of images earlier in time to formulate our multi-scale representation. We
Fig. 5  Our aerial experiments constitute long-range journeys through radically different environments undergoing significant appearance change. Top: each individual navigation trial contains 500 images, randomly selected from images captured at any position in either of the two routes which span rural and highly urbanized environments. Bottom: we show four different image pairs which were successfully matched using our system, even across significant appearance variations.

Table 2  Ground dataset specifications and parameters

| Specification                          | Value                  |
|---------------------------------------|------------------------|
| GPS tolerance for AUC graphs          | 5 frames (100 m)       |
| Total number of reference set images  | 8280                   |
| Number of query images per trial      | 500                    |
| Number of navigation trials           | 20                     |
| Real-world distance between successive images | Approx. 20 m |
| Image resize for SAD                   | 32 by 64 pixels        |
| Image resize for Gist                  | 128 by 128 pixels      |
| Image resize for NetVLAD               | 224 by 224 pixels      |

5 Results

In this section, we provide results from four different types of experiments, each testing a different aspect of the multi-scaled localization approach. Our results begin by considering the aerial dataset in the first instance and then provide further analysis on the ground dataset to demonstrate the versatility of the suggested procedures. In our first experiment on the aerial dataset, we observe the localization performance of each scale size in isolation, and then for our second experiment we combine increasing numbers of scales together. We also provide results which show that the multi-scale representation is particularly suited to improving the positional accuracy of a localization system. The results continue with an experiment to observe the change in localization performance when the absolute scale sizes are increased and then conclude with a similar study on the ground dataset. Localization performance is measured using the AUC (area under the...
Our forward-facing, ground-vehicle experiments consist of navigation trials within a very long journey through mixed rural and urban scenery across different seasons. We use a segment of the Nordland dataset, with the train route shown in the above extract from Google Earth (Google 2020). In our experiments, we attempt to localize our position within this entire route during winter, using a collection of images recorded during summer as our reference. ©2020, Google

precision–recall (precision–recall curve) metric. To generate an AUC graph, we produce multiple value pairs of precision and recall by varying the quality threshold in our localization method. This threshold is the inflection point at which a robot would assess whether it is in a novel location or whether it is in a location it has been before (i.e., perform localization).

5.1 Aerial dataset experiments

5.1.1 How does localization performance vary with different place size representations?

In Fig. 7, we display single-scale localization results using three visual place representation and recognition techniques, NetVLAD, Gist and SAD, that represent the range of technical approaches used in current research. Each bar represents the mean AUC score for that particular scale size, with each AUC data-point generated from a random navigation trial containing 500 query images. In total, we ran 20 navigation trials per scale size and the error bars show the spread of the AUC scores up to one standard deviation from the mean. Each scale reflects a real-world square area with side dimensions, from smallest (scale 1) to largest (scale 9), of 56 m, 79 m, 112 m, 158 m, 224 m, 316 m, 448 m, 633 m and 896 m. We use a scale ratio of \(\sqrt{2}\), reflecting the scale ratios found in the natural system (Stensola et al. 2012).

The results indicate that performance varies at different place scales and between different techniques. The mean AUC trends upwards from the smallest scale toward the largest scale, excepting a small drop in localization performance at the very largest scales for NetVLAD and Gist. These trends can be attributed to improved navigation capabilities when more visual information is provided; however, if too large a scale is used, the localization ability reduces (see Sect. 6 for further discussion). Very small scales are difficult for SAD due to viewpoint variations between the two traverses—as a small scale is a small field-of-view, slight variations in the navigation route between traverses are particularly impactful, whereas Gist and NetVLAD have some intrinsic compatibility to small viewpoint changes, as they use either learnt features or hand-crafted filters rather than simple pixel-wise comparisons.

The best scale for each image processing method is as follows: scale 6 (316 by 316 m) for NetVLAD, scale 7 (448 by 448 m) for Gist and scale 7 (448 by 448 m) using SAD.
5.1.2 How does localization performance vary with the granularity and number of map scales?

In this experiment, we combine several scales together, using the methodology detailed in Sect. 3. When we combine two scales, to isolate the effect of the number of scales, rather than the range captured by the scales, we use the largest scale combined with the smallest scale. For additional scales, we increase the granularity of the scale range by adding more scales within the span of the largest and smallest scales. Our selected scale order therefore keeps the total range of place representations constant. Our multi-scale combinations are shown in Table 3.

The three graphs shown in Fig. 8 present the mean multi-scale localization performance for each of the three different image processing methods using a universal localization tolerance of 50 m from the ground-truth location. For NetVLAD and Gist, using multiple scales improves localization over a single scale, with diminishing returns beyond three scales. We find that SAD is scale-independent at larger scales and is able to localize almost as well at the largest scale as with multiple scales. This is because SAD performs direct pixel comparisons, and as our aerial images are relatively similar in appearance, the simplest algorithm works as well as more complex algorithms like NetVLAD and Gist. However, it is interesting to observe that the two-scale trial still performs surprisingly well, given it is combining a well-performing scale size (the largest scale) with a very poor performing scale (the smallest scale). This result can be attributed to properties of our combination algorithm, such as applying periodicity to small scale sizes.

For all three methods, using three scales provides near peak performance, without the extra computational overheads of having more scales. Interestingly, this result suggests an optimal spacing ratio of 4, rather than the \( \sqrt{2} \) found in the biological system. Intuitively, this larger scale ratio suggests a larger degree of information redundancy between similar scales.

We also perform a statistical analysis to quantitatively verify that simultaneously using multiple scales is an improvement over a single scale, across all three image processing methods. We use a \( t \) test with a significance of \( p < 0.05 \) and 38 degrees of freedom. We show our \( t \) test scores in Table 4, comparing a single scale (we use the largest scale) to each combination of multiple scales. All scores except the two-scale trial using Gist and SAD are greater than the critical \( t \) value of 2.1009, for the specified number of degrees of freedom and \( p < 0.05 \). Our statistical analysis verifies our hypothesis that using multiple spatial scales improves localization performance compared to a single scale; in particular, using three scales provides an ideal compromise between localization performance and computation requirements. Please refer to Sect. 3.2 for further details of our statistical test methodology.

5.1.3 How do multiple map scales affect localization precision?

Following on from our results in Fig. 8, we investigate the relationship between recall performance and precision requirements, that is, the allowable distance error from the
Fig. 8 Localization performance using varying numbers of spatial scales. Mean AUC score for the largest single scale (scale 9) versus increasing numbers of multiple scales, using a NetVLAD, b Gist and c sum of absolute differences.

Table 4 $T$ test scores for NetVLAD, Gist and SAD, comparing single-scale (the largest scale) localization performance to using a varying number of scales.

| Number of Scales | NetVLAD | Gist | SAD  |
|------------------|---------|------|------|
| 2                | 4.55    | -2.18 | -6.67 |
| 3                | 32.42   | 27.78 | 3.03  |
| 4                | 38.77   | 33.86 | 4.80  |
| 5                | 41.41   | 34.19 | 7.34  |
| 6                | 43.00   | 36.73 | 3.77  |
| 7                | 41.94   | 27.68 | 6.91  |
| 8                | 43.32   | 35.57 | 5.65  |
| 9                | 43.37   | 35.12 | 7.39  |

Table 5 Absolute scale sizes with increasing sizes.

| Number of combined scales | Size ID | Real-world sizes |
|---------------------------|---------|------------------|
| 8                         | 1       | 56, 79, 112, 158, 224, 316, 448, 633 m |
| 8                         | 2       | 59, 84, 119, 167, 237, 335, 475, 671 m |
| 8                         | 3       | 63, 89, 126, 178, 252, 355, 503, 711 m |
| 8                         | 4       | 67, 94, 133, 188, 267, 376, 534, 754 m |
| 8                         | 5       | 71, 100, 141, 199, 283, 399, 566, 799 m |
| 8                         | 6       | 75, 106, 150, 211, 300, 423, 600, 847 m |
| 8                         | 7       | 79, 112, 158, 224, 316, 448, 633, 896 m |

ground-truth location. In Fig. 9, the recall rate is the number of correct localizations, assuming that every query scene has a matched image in the reference set (for our experiments this holds true). We define a correct localization if the GPS distance between the current scene and the matched scene is less than a threshold, which is varied as per the x-axis of Fig. 9. We compare the recall rate using the combination of all nine scales against each and every single (baseline) scale.

When high localization precision is required, using multiple spatial scales is the optimal approach, with particular improvement for NetVLAD. For Gist and NetVLAD, the worst performing single scale at a 10-m tolerance is the

Fig. 9 Localization recall performance when varying the required localization precision. “With Scales” is using all 9 scales simultaneously, compared to all individual baseline scales. Using multiple scales improves the recall rate across the board, but particularly when a high localization precision is required (such as a GPS distance of 10 m from an exact matching location). The minimum GPS accuracy is 10 m, which is why there are no recall rate data between 0 and 10 m on the graphs.
largest scale, which is expected since the largest scale observes a visual region significantly larger than the localization tolerance. Additionally, NetVLAD and Gist have specific optimal scale size properties, whether that be due to training (NetVLAD), or configuration of spectral filters (Gist). SAD experiences less deficiencies at small localization tolerances due to its direct pixel matching, but conversely cannot match the absolute localization performance of either NetVLAD or Gist.

5.1.4 How does the absolute scale size affect multi-scale localization performance?

In our final experiment, we consider the ideal absolute scale size; does increasing the real-world size of the smallest scale and all other larger scales change the performance characteristics of a multi-scale system? We vary the size of the smallest scale and larger scales while maintaining both the total scale range ratio and the inter-scale ratio of $\sqrt{2}$. The scale sizes for each variation are shown in Table 5 (expressed in terms of the real-world place width in meters). We combine eight scales for different initial scale sizes, across the three image processing methods. These results are shown in Fig. 10.

Figure 10 reveals a dissimilarity in the impact of different absolute scale sizes across the different image processing methods. To quantitatively evaluate the variations across different scale sizes, we perform a statistical $t$ test between size one and all other sizes, using the mean and standard deviation from 20 navigation trials. Our alternate hypothesis is that the absolute scale size does change the localization success rate, even when the number of scales is constant. At 38 degrees of freedom and a significance of $p < 0.05$, the corresponding critical $t$ value is 2.0244.

For NetVLAD, the $t$ test produced the values shown in Table 6 from size two to size seven. The smallest absolute scale size is potentially the best scale size; however, given the lack of a definitive trend over the scale range, it is more likely that NetVLAD is independent of the absolute scale size in a multi-scale system.

The absolute scale size does matter when Gist and SAD are examined. For Gist, the larger the size of the smallest scale, the better the localization performance. Comparing scale size one and scale size seven, a $t$ value of 8.73 is larger than a critical $t$ value of 2.021 given 38 degrees of freedom and a significance of $p < 0.05$.

For SAD, we again run a $t$ test between scale one and the other absolute scale sizes. The low variation between navigation trials for SAD reveals the same increasing trend as Gist; that is, localization performance improves as the absolute scale size is increased. For all absolute scale sizes except scale four, our calculated $t$ value exceeds the critical $t$ value of 2.021 at a 5% significance level. These results will be discussed further in Sect. 6.3.

5.2 Ground dataset

For our ground dataset (Nordland), we show results for different individual scale sizes and results for different numbers of multiple scales. As described in Sect. 3.2, a scale is a down-sampled feature vector from a sequential concatenation of image features. With this method, each scale size has a uniform feature descriptor dimensionality; however, each successive scale includes visual features from increasingly large visual regions. Like for the aerial dataset, our multi-scale combinations maintain a fixed span between the largest and smallest scales, and we increase the granularity of the scale spacing when we add additional scales. Our multi-scale combinations are shown in Table 7.

It is worth noting that this dataset is significantly more challenging to localize in, compared to the Nearmap aerial dataset. Two reasons exist: the increased size of the navi-

Fig. 10 Mean AUC localization performance using different combinations of eight scales. Absolute scale size increases from size one to size seven. We evaluate three different image processing methods and perform 20 random navigation trials.
Table 7 Combinations of multiple scales used in our experiments

| Number of combined scales | Scale IDs | Real-world sizes (sequence length in frames) |
|---------------------------|-----------|---------------------------------------------|
| 2                         | 1, 9      | 1, 23                                       |
| 3                         | 1, 5, 9   | 1, 2, 6, 23                                 |
| 4                         | 1, 4, 6, 9| 1, 4, 8, 23                                 |
| 5                         | 1, 3, 5, 7, 9 | 1, 3, 6, 11, 23                          |
| 6                         | 1, 2, 4, 6, 8, 9 | 1, 2, 4, 8, 16, 23                  |
| 7                         | 1, 3, 4, 5, 6, 7, 9 | 1, 3, 4, 6, 8, 11, 23            |
| 8                         | 1, 2, 3, 4, 6, 7, 8, 9 | 1, 2, 3, 4, 8, 11, 16, 23      |
| 9                         | 1, 2, 3, 4, 5, 6, 7, 8, 9 | All sizes                           |

Table 8 Multi-scale statistical analysis on Nordland

| Number of scales | NetVLAD | Gist | SAD |
|------------------|---------|------|-----|
| 2                | 3.13    | -13.74 | -6.67 |
| 3                | 20.98   | 1.35 | 3.03 |
| 4                | 13.51   | -9.66 | 4.80 |
| 5                | 29.07   | 5.30 | 7.34 |
| 6                | 16.26   | -5.87 | 3.77 |
| 7                | 24.26   | -2.16 | 6.91 |
| 8                | 25.72   | 2.29 | 5.65 |
| 9                | 25.60   | -0.60 | 7.39 |

The results for single scale sizes can be seen in Fig. 11. For Gist and SAD, the largest scale produces the highest AUC score, as this scale has accumulated the largest amount of visual information. The drop-off in localization that was observed using Gist for our aerial dataset does not occur. This is likely due to the fact that our ground algorithm does not resize the images to produce multiple scales; thus, there is no performance loss due to potentially unfavorable spatial frequencies in the images. NetVLAD performs poorly at all scales, partly due to the training environment of NetVLAD (Pittsburgh) being very different to deployment. Dimensionality reduction also has some effect on NetVLAD. It is important to note that the absolute performance of any particular technique here is not as important as the variation of performance across different scales and when combining multiple scales.

Fig. 11 Single-scale localization results on the Nordland dataset. We display the mean AUC for different sized place scales, from the smallest scale (scale 1) through to the largest scale (scale 9), with error bars denoting the spread of results over the localization trials. The three image processing methods exhibit similar performance characteristics over scale size, although NetVLAD has an absolute localization success rate much lower than either Gist or SAD

Fig. 12 Localization performance using varying numbers of spatial scales on the Nordland dataset. Mean AUC score for the largest single scale (scale 9) versus increasing numbers of multiple scales, using a NetVLAD, b Gist and c sum of absolute differences

![Image](https://via.placeholder.com/150)
5.2.2 How does localization performance vary with the granularity and number of map scales?

When multiple scales are used for localizing using the ground algorithm, a similar trend is observed as with the aerial dataset (Fig. 12). Localization performance (expressed using the AUC metric) improves up to three scales, after which additional scales become increasingly redundant. Our AUC scores are less consistent over the range of scales for NetVLAD and Gist, because of anomalies in the single-scale performance at scale sizes 4, 6 and 8 (sequence lengths 4, 8 and 16). These anomalies are artifacts of our sampling approach to dimensionality reduction; NetVLAD in particular has a high inter-vector performance variation. In other words, certain elements of the feature vector are better for localization in the chosen dataset than others, which in itself stems from specific feature maps having differing performance characteristics (Hausler et al. 2020). The combination of our dimensionality reduction plus utilizing multiple scales enables improved localization with a cheap computational cost, although the localization improvement is marginal for Gist. Conversely, our results also reveal that Gist and SAD, when provided features from a temporal sequence, are much more efficient at localizing than NetVLAD when all three methods are normalized to an equal feature vector length. Again we run a statistical analysis, comparing the largest single scale to differing numbers of multiple scales (Table 8). The analysis shows that combining five scales (sequence lengths 1, 3, 6, 11 and 23) provides the best improvement to localization compared to a single scale.

5.2.3 How do multiple map scales affect localization precision?

We again display a figure showing the recall rate over ground-truth tolerance (Fig. 13). We define the tolerance as the frame distance between the current scene and the matched scene, since the dataset is frame-aligned. We investigate the recall rate using all nine scales compared to each individual baseline scale. Similar trends are present as in the aerial experiment; NetVLAD continues to have the greatest improvement using multiple scales. All three image processing methods have difficulty achieving a high recall rate, due to the difficulty of the dataset. As discussed earlier, NetVLAD experiences major difficulties due to differences between the deployment environment and the environment the network was originally trained on. A final observation of note is that the multi-scale approach has the greatest recall rate using Gist when exact frame-aligned matching is required (a tolerance of zero frames).

5.3 Homogeneous multi-scale mapping in action

In Fig. 14, we show several examples of place matches using our multi-scale approach (on the aerial dataset). We include three example query locations, with two successful place matches and one failure case. In the top row, we display the different scale place views as well as the overall view of the query location (at far right), while on the bottom row, we visualize the best matching place at each individual scale, with the overall place match chosen by our multi-scale sys-
Fig. 14 Three example place matches, displaying both the multi-scale matched scene and the matched location hypothesized by each single scale. Distances listed are the matching errors: the distance between the matched reference location and the correct query location. If this distance is less than the ground-truth tolerance (50 m), the text is colored green. 

a Due to significant appearance change, none except one of the best single-scale location matches are correct, but the fusion of all scales results in the correct overall location match. 

b A dense urban environment provides a somewhat easier localization challenge, with single scales 5 and 9 being correct, as well as the overall multi-scale estimate being correct. 

c A failure example: While scale 8 is close to finding the correct location, the severe perceptual aliasing across the remaining scales produced a false multi-scale match (color figure online).

5.4 Effect of periodicity on the results

We conclude our results by providing an experimental analysis of the impact of periodic normalization on the localization success rate, measured using the AUC metric. Figure 15 considers the AUC result with all nine scales, with and without the addition of periodic normalization. In the “without” case, each scale size is still separately globally normalized.

The results show that the addition of local periodic normalization is essential for ground-based sensory input (Nordland), while unnecessary for aerial sensory input (Nearmap). This can be explained by the difference in dataset structure. The reference image set for the Nearmap aerial dataset is not an ordered list—subsequent images are not necessarily adjacent in the real world, since we combine two very dissimilar real-world locations (dense urban and rural) to produce our dataset. Local periodic normalization is best used when the reference set is spatially ordered, as is the case for any robot or animal moving on the ground-plane, such as vehicles or trains in this particular case. Local normalization then finds the best match out of a set of images from a similar real-world location. Interestingly, NetVLAD gains no benefits from periodic normalization in either dataset. It can be speculated that the localization performance with NetVLAD on Nordland is deficient to the extent that finding local best matches cannot overcome the fundamentally poor recognition capabilities in this environment.

6 Discussion and conclusion

Inspired by the homogeneous, multi-scale mapping systems found in the mammalian brain, we have performed a series of investigations into several of the key characteristics of these natural mapping systems using three place recogni-
6.1 Multiple benefits of multi-scale fusion

Using multiple map scales universally improves localization performance, and for the most part, regardless of the type of place representation or comparison technique used. We discovered that our techniques are particularly applicable to NetVLAD, which the authors hypothesize reveals an interesting complementarity between bio-inspired neural networks and bio-inspired navigation algorithms.

The localization performance improvements from multiple scales manifest themselves in several functionally relevant ways: Absolute AUC performance improves, representing the overall performance across a range of different operating zones, but performance at any particular localization precision level improves even more drastically. We found that three scales with a large scale ratio provide improved localization performance and further increases to scale granularity are not essential. The performance gains from combining multiple scales can be explained by considering the utility of different sized scales. The representation of a place at a large scale inherently has access to more information, even though we normalize all place representations during usage to a standard template size. This extra information reduces the likelihood of place match aliasing, but comes at a cost: The ability of a localization system to localize at sub-ten meter accuracy decreases when provided visual information that spans a broad area of the environment. When large scale and small scale are combined, the place recognition system can utilize the best of both scale sizes. The single-scale results are particularly interesting in this regard: there is great variation in performance at different single map scales, but “bigger” is not always better as shown in both Figs. 7 and 9. Additionally, for NetVLAD and Gist, the large and small scales can provide different visual feature patterns which can be complementary for these particular algorithms (as observed in both the aerial and ground datasets).

6.2 Comparing different processing methods with our approach

The relevance of our proposed approach varies quite significantly with respect to the image processing method used. With a single scale, on both datasets, Gist and SAD localize significantly better than NetVLAD. This is likely due to the pre-trained aspect of NetVLAD, which was not trained on aerial images nor long distance traverses in remote terrain (Nordland). However, combining multiple scales is particularly efficient for NetVLAD, which suggests that there may be an underlying benefit to fusing different spatial scales at the input stage to a neural network. Future work will investigate this hypothesis in further detail. For both datasets, Gist and SAD (which both have no training component) exhibit a trend of increasing localization performance as the place representation scale increases, whether that be due to a larger field-of-view or a longer sequence of images.

6.3 Impact of varying the absolute scale between image processing methods

Our final experiment for the aerial dataset (see Fig. 10) investigated the effect of the absolute scale sizes used in a multi-scale system. Variations in behavior were observed: Using NetVLAD, localization was not affected by varying the absolute scale size, while Gist and SAD displayed an increase in localization performance as the absolute scale size increased. This behavior matches the trends observed in Fig. 7, where the larger scales are nearly always better than the smaller scales. Practically, these results suggest that
the choice of absolute scale size for a multi-scale mapping approach will assume varying levels of importance depending on the type of place representation technique utilized, with NetVLAD being the most scale-independent and hence the most “out-of-the-box” ready without the need for scale calibration.

### 6.4 Computation requirements analysis

In Tables 9 and 10, we present a computation time analysis of our multi-scale approach across the three place recognition methods on both datasets, comparing a single scale, three scales and all nine scales. We present average computation times per place, as well as feature extraction and place matching process components.

The computation times are calculated using MATLAB 2018b on an i7-7700K CPU, although these values are approximate as the computation times had some variation across navigation trials. Of the three methods, Sum of Absolute Differences would be the best suited to operation on an embedded device on a mobile autonomous vehicle, particularly if fewer scales are used. However, no particular optimization has been performed in this work: It is likely that these computation times could be significantly reduced with commonly used optimization techniques. The computational efficiency (increase in AUC with respect to increase in compute time) of multiple scales reduces with each additional scale. It is also worth noting that the two methods (aerial and ground) have different calculations with very different time constraints. For example, the largest computational bottleneck on Nordland is the sequence generation component within the feature extraction step, an operation which scales with respect to the size of the original feature vector. The matching step on Nordland is very fast, since we use feature vectors with dimension 400 for all three techniques.

### 6.5 Complex spatial navigation

In a classical view of autonomous navigation, the perceived world is observed up to the field-of-view of the sensor used. This view contains information about both nearby and distant visual landmarks. However, the uniqueness of each landmark to the current position in the world is dependent on the distance between the current position and these landmarks. Proximal landmarks are good for positioning as they are correlated with the position of the autonomous agent. Distal landmarks can be observed over an extended period of movement, making them less suited to precise localization, but more informative in estimating the current bearing. In this work, we intentionally separate the visual landmarks (from a birds-eye view, using crop and zoom functions) by distance (within the plane of the scene) to the observer, then localize each visual scale separately and merge these separate place estimates. To clarify, this distance is the inter-image distance, which, in real-world distance, can be thought of as the opposite side of a right-angled triangle where the angle of this triangle is the incident view angle from the observer toward the target landmark. Such an approach is trivial for laser-based robotic platforms with an inertial measurement unit; however, for just a visual sensor this methodology is a more complex representation of the visually observable world. Recent research discovered a cell type called “object-vector cells” in the rodent medial entorhinal cortex (Høydal et al. 2019). Individual cells fire when rodents are at fixed distances and directions from landmarks in their environment. Therefore, the evidence suggests that the mammalian brain is also intentionally separating visually observable landmarks by spatial proximity, before integrating this information in grid and place cells. Such spatially complex world representations are also been observed in other species, such as...
hummingbirds (Hurly et al. 2014). Our forward-facing experiment loosely separates distal and proximal landmarks using sequences of images; however, as forward-facing images are non-planar, future work will need to estimate the inter-image depth (using stereo vision and scene geometry) in order to accurately separate these types of landmarks in the observed scene.

6.6 Future directions

The results from this research show a clear motivation for implementing a homogeneous, multi-scale approach to mapping and localization regardless of the particular place representation, or place comparison methods being used, or type of navigation (aerial versus ground). Our results suggest a range of further studies that may reveal additional performance improvement opportunities in the field of robotics and provide further stimulation for biological studies with animals in large-scale navigation experiments.

Our experimental results suggest that peak performance is achieved with a much larger scale ratio between adjacent map scales (below which there are diminishing performance gains) than has been found in biological systems to date. We investigated the computational performance over number of scales; however, it would also be interesting to further investigate at what scale ratio diminishing performance returns are encountered when systematically varying the place representation, the place comparison technique and the environment type and conditions. Such analysis will determine whether the optimal scale ratios vary depending on these factors: It may also be possible to closely replicate the properties of these systems for specific species like rodents, to see whether a similar ratio is observed to those seen in the brain.

From a practical perspective, we have used a range of approaches that represent the range of state-of-the-art place representations used in robotic localization and demonstrated improvements resulting from the multi-scale approach. This is particularly true when our multi-scale approach is used in conjunction with a learnt representation like NetVLAD. Future work could investigate a range of additional optimizations and practical improvements to this work. The most valuable improvement would be to expand this work to a fully two-dimensional global localization solution. Such an approach would consider a place in the metric sense, with image representations captured in conjunction with motion and pose estimates, utilizing joint motion and observation techniques such as particle filters (Grissetti et al. 2007) or pose-graph optimization (Lu and Milios 1997). Future work could also investigate novel and alternate two-dimensional localization solutions by further investigating biological navigation. Secondly, there are a range of promising optimizations that could enhance the performance or computational requirements of the systems presented in this paper. While our multiple map scale localization estimates are processed independently before being fused together, it should be possible to implement a hierarchical approach where matches at one scale are used to filter the candidate matches at subsequent scales, significantly reducing compute requirements, and even potentially improving localization performance. It would also be worthwhile to further investigate the computational cost of multi-scale systems, with respect to both the energy cost (compute time) and representational cost (amount of memory or neurons required).

Finally, in the place-based, rather than frame-based, approach presented here, there are a myriad of ways in which a “place” could be represented, especially when considering the range of ways in which a place can be observed by different sensing modalities (compare cameras with LIDAR and radar) on different robotic platforms and in different environmental domains on or under the ground or sea, and in the air. New neural recordings from bats (Geva-Sagiv et al. 2015) and rodents (Casali et al. 2019) navigating in large, three-dimensional environments may yield new inspiration for how these place representations can be implemented in a robotic system. These differences in sensors, “platform” and operating environment are also reflected in the natural kingdom: further investigation into possible mechanisms for representing place in a multi-scale mapping and localization framework may yield insights that both aid robotics and our understanding of these natural systems.

References

Arandjelovic R, Gronat P, Torii A, Pajdla T, Sivic J (2018) NetVLAD: CNN architecture for weakly supervised place recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1437–1451. https://doi.org/10.1109/TPAMI.2017.2711011. arXiv:1511.07247

Arleo A, Gerstner W (2000) Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. Biol Cybern. https://doi.org/10.1007/s004220000017

Banino A, Barry C, Uria B, Blundell C, Lillicrap T, Mirowski P, Pritzel A, Chadwick MJ, Degris T, Modayili J, Wayne G, Soyer H, Viola F, Zhang B, Goroshin R, Rabinowitz N, Pascanu R, Bastie C, Petersen S, Sadik A, Gaffney S, King H, Kavucuoglu K, Hassabis D, Hadsell R, Kumaran D (2018) Vector-based navigation using grid-like representations in artificial agents. Nature 557(7705):429–433. https://doi.org/10.1038/s41586-018-0102-6

Bosse M, Newman P, Leonard J, Soika M, Feiten W, Teller S (2003) An Atlas framework for scalable mapping. In: 2003 IEEE international conference on robotics and automation (Cat. No.03CH37422). IEEE, pp 1899–1906. https://doi.org/10.1109/ROBOT.2003.1241872

Browning B (2000) Biologically plausible spatial navigation for a mobile robot. Computer Science and Electrical Engineering

Brun VH, Solstad T, Kjelstrup KB, Fyn M, Witter MP, Moser EI, Moser MB (2008) Progressive increase in grid scale from dorsal to ventral medial entorhinal cortex. Hippocampus 18(12):1200–1212. https://doi.org/10.1002/hipo.20504
Burak Y, Fiete IR (2009) Accurate path integration in continuous attractor network models of grid cells. PLoS Comput Biol 5(2):e1000291. https://doi.org/10.1371/journal.pcbi.1000291
Casali G, Bush D, Jeffery K (2019) Altered neural odometry in the vertical dimension. Proc Nat Acad Sci USA 116(10):4631. https://doi.org/10.1073/pnas.1811867116
Chen Z, Lowry S, Jacobson A, Hasselmo ME, Milford M (2015) Bio-inspired homogeneous multi-scale place recognition. Neural Netw 72:48–61. https://doi.org/10.1016/j.neunet.2015.10.002
Cummins M, Newman P (2009) Highly scalable appearance-only SLAM. In: Robotics science and systems V
Erdem U, Hasselmo M (2014) A biologically inspired hierarchical goal directed navigation model. J Physiol Paris 108(1):28–37. https://doi.org/10.1016/j.jphysparis.2013.07.002
Furgale P, Barfoot TD (2010) Visual teach and repeat for long-range rover autonomy. J Field Robot 27(5):534–560. https://doi.org/10.1002/rob.20342
Geva-Sagiv M, Las L, Yovel Y, Ulanovsky N (2015) Spatial cognition in bats and rats: from sensory acquisition to multiscale maps and navigation. Nat Rev Neurosci 16(2):94–108. https://doi.org/10.1038/nrn3888
Google (2020) Google maps. https://www.google.com/maps. Accessed 21 Feb 2020
Grisetti G, Tiefaldi GD, Stachniss C, Burgard W, Nardi D (2007) Fast and accurate SLAM with Rao-Blackwellized particle filters. Robot Auton Syst. https://doi.org/10.1016/j.robot.2006.06.007
Hafting T, Fyhn M, Molden S, Moser MB, Moser EI (2005) Microstructure of a spatial map in the entorhinal cortex. Nature 436(7052):801–806. https://doi.org/10.1038/nature03721
Haasler S, Jacobson A, Milford MJ (2019) Multi-process fusion: visual place recognition using multiple image processing methods. IEEE Robot Autom Lett. https://doi.org/10.1109/LRA.2019.2898427
Høydal ØA, Skytøen ER, Andersson SO, Moser MB, Moser EI (2019) Spatial cognition in bats and rats: from sensory acquisition to multiscale maps and navigation for sunny summer days and stormy winter nights. In: Proceedings—IEEE international conference on robotics and automation. IEEE, pp 1643–1649. https://doi.org/10.1109/ICRA.2019.8246263
Hurly TA, Fox TA, Zwueste DM, Healy SD (2014) Wild hummingbirds rely on landmarks not geometry when learning an array of flowers. Anim Cognit 17(5):1157–1165. https://doi.org/10.1007/s10071-014-0748-x
Jacobson A, Scheirer W, Milford M (2017) Déjà vu: scalable place recognition using mutually supportive feature frequencies. In: IEEE international conference on intelligent robots and systems. https://doi.org/10.1109/IROS.2017.8206580
Jacobson A, Chen Z, Milford M (2018) Leveraging variable sensor spatial acuity with a homogeneous, multi-scale place recognition framework. Biol Cybern 112(3):209–225. https://doi.org/10.1007/s00422-017-0745-7
Kriechmar JL, Seth AK, Nitz DA, Fleischer JG, Edelman GM (2005) Spatial navigation and causal analysis in a brain-based device modeling cortical-hippocampal interactions. Neuroinformatics. https://doi.org/10.1007/s13357-003-0519-7
Kuipers B (2007) An intellectual history of the spatial semantic hierarchy. In: Jeffries ME, Yeap WK (eds) Robotics and cognitive approaches to spatial mapping. Springer, Berlin, pp 243–264. https://doi.org/10.1007/978-3-540-75388-9_15
Lever C, Burton S, Jeewajee A, O’Keefe J, Burgess N (2009) Boundary vector cells in the subiculum of the hippocampal formation. J Neurosci 29(31):9771–9777. https://doi.org/10.1523/jneurosci.1319-09.2009
Lowry S, Sunderhauf N, Newman P, Leonard JJ, Cox D, Corke P, Milford MJ (2016) Visual place recognition: a survey. IEEE Trans Robot 32(1):1–19. https://doi.org/10.1109/TRO.2015.2496823. arXiv:1612.00593
Lu F, Milios E (1997) Robot pose estimation in unknown environments by matching 2d range scans. J Intell Robot Syst Theory Appl 18(3):249–275. https://doi.org/10.1023/A:1007957421070
Milford MJ, Wyeth GF (2008) Mapping a suburb with a single camera using a biologically inspired SLAM system. IEEE Trans Robot 24(5):1038–1053. https://doi.org/10.1109/TRO.2008.2004520
Milford M, Wyeth G (2010) Persistent navigation and mapping using a biologically inspired slam system. Int J Robot Res 29(9):1131–1153. https://doi.org/10.1177/0278364909340592
Mount J, Dawes L, Milford MJ (2019) Automatic coverage selection for surface-based visual localization. IEEE Robot Autom Lett 4(4):3900–3907. https://doi.org/10.1109/LRA.2019.2928259
Mur-Artal R, Montiel JM, Tardos JD (2015) ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Trans Robot 31(5):1147–1163. https://doi.org/10.1109/TRO.2015.2463671. arXiv:1502.00956
Naseer T, Burgard W, Stachniss C (2018) Robust visual localization across seasons. IEEE Trans Robot 34(2):289–302. https://doi.org/10.1109/TRO.2017.2788045
Nearmaps (2019) Nearmaps: aerial maps, high resolution aerial imagery. https://www.nearmap.com/au/en. Accessed 01 Aug 2019
O’Keefe J, Conway D (1978) Hippocampal place units in the freely moving rat: why they fire where they fire. Exp Brain Res 31(4):573–590. https://doi.org/10.1007/BF00239813
Oliva A, Torralba A (2006) Building the gist of a scene: the role of global image features in recognition. Prog Brain Res 155B:23–36. https://doi.org/10.1016/S0079-6123(06)55002-2
Sargolini F, Fyn M, Hafting T, McNaughton BL, Witter MP, Moser MB, Moser EI (2006) Conjunctive representation of position, direction, and velocity in entorhinal cortex. Science 312(5774):758–762. https://doi.org/10.1126/science.1125572
Schoenenberger P, O’Neill J, Csicsvari J (2016) Activity-dependent plasticity of hippocampal place maps. Nat Commun 7:11824. https://doi.org/10.1038/ncomms11824
Sherry DF, Grella SL, Guigueno MF, White DJ, Marrone DF (2017) Are there place cells in the avian hippocampus? Brain Behav Evol 90(1):73–80. https://doi.org/10.1159/000477085
Sreenivasan S, Fiete I (2011) Grid cells generate an analog error-correcting code for singularly precise neural computation. Nat Neurosci. https://doi.org/10.1038/nmn.2901
Stemmler M, Mathis A, Herz AV (2015) Neuroscience: connecting multiple spatial scales to decode the population activity of grid cells. Sci Adv 1(11):1–12. https://doi.org/10.1126/sciadv.1500816
S Sünderhauf N, Neubert P, Protzel P (2013) Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons. In: Workshop on long-term autonomy at ICRA 2013. IEEE.
Sünderhauf N, Shirazi S, Dayoub F, Upcroft B, Milford M (2015) On the performance of ConvNet features for place recognition. In: IEEE international conference on intelligent robots and systems, vol 2015-Decem. IEEE, pp 4297–4304. https://doi.org/10.1109/IROS.2015.7353986. arXiv:1501.04158

Taube JS, Muller RU, Ranck JB (1990) Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. J Neurosci 10(2):420–35. https://doi.org/10.1523/JNEUROSCI.10-02-00420.1990

Ulanovsky N (2015) Three-dimensional head-direction coding in the bat brain. Nature 517(7533):159–164. https://doi.org/10.1038/nature14031

Yu L, Jacobson A, Milford M (2017) Rhythmic representations: learning periodic patterns for scalable place recognition at a sub-linear storage cost. https://doi.org/10.1109/LRA.2018.2792144. arXiv:1712.07315

Yu F, Shang J, Hu Y, Milford M (2019) NeuroSLAM: a brain-inspired SLAM system for 3D environments. Biol Cybern. https://doi.org/10.1007/s00422-019-00806-9

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.