Leveraging variable sensor spatial acuity with a homogeneous, multi-scale place recognition framework

Adam Jacobson1 · Zetao Chen1 · Michael Milford1

Received: 31 August 2016 / Accepted: 21 December 2017 / Published online: 20 January 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract
Most robot navigation systems perform place recognition using a single-sensor modality and one, or at most two heterogeneous map scales. In contrast, mammals perform navigation by combining sensing from a wide variety of modalities including vision, auditory, olfactory and tactile senses with a multi-scale, homogeneous neural map of the environment. In this paper, we develop a multi-scale, multi-sensor system for mapping and place recognition that combines spatial localization hypotheses at different spatial scales from multiple different sensors to calculate an overall place recognition estimate. We evaluate the system’s performance over three repeated 1.5-km day and night journeys across a university campus spanning outdoor and multi-level indoor environments, incorporating camera, WiFi and barometric sensory information. The system outperforms a conventional camera-only localization system, with the results demonstrating not only how combining multiple sensing modalities together improves performance, but also how combining these sensing modalities over multiple scales further improves performance over a single-scale approach. The multi-scale mapping framework enables us to analyze the naturally varying spatial acuity of different sensing modalities, revealing how the multi-scale approach captures each sensing modality at its optimal operation point where a single-scale approach does not, and enables us to then weight sensor contributions at different scales based on their utility for place recognition at that scale.

Keywords Place recognition · Place cells · Sensor fusion · Robotics

1 Introduction

Studies of rodent physiology and neurology have revealed that rats, and likely many mammals, build a homogeneous multi-scale neural map of the environment using multiple sensing modalities to enable localization and navigation. At the heart of this neural representation is the “place cell,” individual neurons which have been found to be spatially sensitive. It has been shown that there is not an individual unique place cell for every unique place in the world; instead, there exists a collection of neurons that activate based on a collection of sensory stimuli, the activation of these cells being strongly correlated to the location of the animal (Fenton et al. 2008; Ravassard et al. 2013; Wiener et al. 2002). The place cells have been shown to represent places of varying size from less than 20 cm to several meters (Solstad et al. 2006; Muller et al. 1987; Jung et al. 1994; Wikenheiser and Redish 2015) and possibly larger. This multi-sensory, multi-scale approach is in direct contrast to conventional robot localization systems, which typically use only one or two sensing modalities and one or at most two spatial map scales (Cummins and Newman 2011; Davison et al. 2007; Andresson et al. 2008; Bosse et al. 2003; Kuipers et al. 2004; Kuipers and Byun 1991). Even when two map scales are used, they typically consist of a local metric map and a global topological map (Bosse et al. 2003; Kuipers et al. 2004). Only recently has the utility of multiple homogeneous map scales been investigated in Chen et al. (2014), by using a single camera sensor to artificially segment data in order to create multiple artificial maps of different scales. Studies fusing multiple sensing modalities for localization have done so at one scale, with no explicit...
framework for encoding and recognizing locations at different spatial scales (Arras et al. 2001).

In this paper, we present five contributions:

1. We develop a localization system for integrating commodity sensors, such as camera, WiFi and barometric sensors, which naturally produce localization results of different spatial resolution (Fig. 1) with a multi-scale homogeneous mapping framework.

2. Our framework encodes the sensory data associated with places at several different scales and then performs recognition over these multiple scales before combining the individual scale place match hypotheses to form a global place match hypothesis.

3. We provide an analysis of the multi-scale framework, capturing each sensor’s contribution to place estimation and evaluating each sensor’s optimal operating scale.

4. We collected and released a new large, multi-sensor place recognition dataset comprising of three 1.5-km day and night traverses of a university campus including outdoor environments and multiple floors of an office building.

5. We provide comparison to a conventional vision-only place recognition system as well as results capturing the contributions of different sensing modalities at different scales.

The paper proceeds as follows: In Sect. 2, we review robotic and animal mapping systems with a focus on multi-scale representations of the world. Section 3 presents our approach describing in detail the place recognition system and the proposed multi-scale sensor fusion technique. In Sect. 4, we present the experimental setup, with the results of multiple levels of evaluation presented in Sect. 5. Section 6 discusses the outcome of the research and areas of future work.

2 Background

In this section, we provide a review of biological navigation processes with a focus on the hippocampal formation in the mammalian brain where the multi-scale neuronal map is encoded. We also provide a review of current research in the areas of place recognition with a focus on systems utilizing cameras or WiFi sensor modalities.

2.1 Biological multi-scale mapping

Rats live and navigate in an incredibly diverse range of environments all over the world. This navigation capability is achieved without genetic specialization to a particular environment, leveraging the same basic neural machinery and sensing modalities. Research into the mapping processes of rats has revealed the importance of the entorhinal–hippocampal formation for rodent navigation within an environment. It is believed that the entorhinal cortex receives and stores spatial and sensory information within the medial entorhinal cortex (MEC) and the lateral entorhinal cortex (LEC), respectively. The entorhinal–hippocampal formation then fuses the information from the MEC and the LEC which is utilized by the hippocampus to perform pattern learning and recognition.

It has been discovered that within the rodent hippocampus, there exist place cells which exhibit spatially correlated firing patterns, believed to enable rodents to encode their location within the world. Unique places within the world appear to be encoded by the firing of a collection of cells which have spatially selective firing patterns, as shown in Fig. 2. The place cells which are active when the animal is in a particular place are the collection of neurons which fire based on multi-
sensory stimuli (Fenton et al. 2008; Ravassard et al. 2013; Wiener et al. 2002).

Place cells are often approximated using Gaussian functions to represent the receptive fields of each neuron, or the spatial area which causes the neurons to fire (Solstad et al. 2006; Muller et al. 1987; Keefe and Burgess 1996). Interestingly, place cells have been shown to encode places with varying spatial resolution, varying in size from less than 20 cm to several meters (Solstad et al. 2006; Muller et al. 1987; Jung et al. 1994; Wikenheiser and Redish 2015) and possibly larger, as shown in Fig. 3.

In this model, the sum of these Gaussian functions is used to model the firing of neurons corresponding to a location within an environment; however, these techniques primarily seek to explain the correlation that exists between the location of rodents and the spatially specific activations of neurons within the hippocampus, not the cause of activation of these neurons within the world. Work presented by Burgess and O’Keefe (1996) demonstrates this model, encoding a rat’s position within a 2D space using

$$\sum_i f_i \overrightarrow{x_i} / \sum_i f_i,$$

where $f_i$ is a radial basis function (or Gaussian function) representing the firing rate of a place cell $i$ centered at $\overrightarrow{x_i}$. This equation only represents the area which is active when the animal is within a specific area within an environment, not what causes this area to become active.

Furthermore, research into the LEC has suggested that rats incorporate sensory information from a number of sensing capabilities including visual, tactile, olfactory (smell), vestibular and auditory stimuli (Hargreaves et al. 2005). This multi-sensory hypothesis is further compounded by behavioral analysis of rats which indicate robust navigational capabilities after sensor failure, injury or the removal of specific sensory cues. It is likely that the internal mapping process derives information from multiple sensory cues as rats are able to reliably navigate both when all sensory modalities are available or after various sensory modalities or environmental cues have been removed (Rossier et al. 2000; Sheppard et al. 2013; Maaswinke and Whishaw 1999), with this behavior also evident in other mammals such as bats (Vanderelst et al. 2016). There is also evidence to suggest that the information between the MEC and LEC is “filtered” to inhibit or excite correct localization using temporal and spatial information from self-motion cues (Hargreaves et al. 2005; Sheppard et al. 2013).

2.2 Place recognition

Individually, cameras and WiFi have been used for localization in many environments. Typically, sensor readings are used to create unique identifiers of places which allow localization. Camera-based place recognition techniques use whole-image (Olivia 2005; Milford and Wyeth 2012) or feature-based techniques (Lowe 1999; Bay et al. 2006) to represent locations. The FABMAP system (Cummins and Newman 2011) leverages a camera to perform mapping over 1000 km, while other camera-only based systems have produced impressive results (Davison et al. 2007; Andreasson et al. 2008; Paz et al. 2008; Konolige and Agrawal 2008; Kawewong et al. 2010). The Atlas system (Bosse et al. 2003) and the hybrid extension to the spatial semantic hierarchy (Kuipers et al. 2004; Kuipers and Byun 1991) have leveraged local metric maps and global topological maps to achieve real-time navigation and map large spaces, but typically using a single or at most two sensing modalities. The discrepancy between the global topological map and local metric maps also complicates integration of place recognition hypotheses formed at a local and global level.

There are a number of techniques which leverage WiFi for localization and mapping including WiFi GraphSLAM (Huang et al. 2011), the distributed particle filter SLAM (DPSLAM) system (Faragher et al. 2012) and others (Weyn 2011; Ferris et al. 2007). These techniques all utilize WiFi fingerprinting, a process similar to camera-based place recognition, where available WiFi devices are detected and the MAC address and the associated signal strength for each device are recorded, producing a unique snapshot of an
environment to enable localization. Berkvens et al. (2014) investigated the combination of WiFi and camera sensing modalities to combat the effect of day/night transitions, but using a conventional single mapping scale framework.

To date, there has been one robotic investigation into the utility of multi-scale maps for place recognition on robots (Chen et al. 2014). The work presented in Chen et al. (2014) was limited in that it was only able to utilize sensory data from a single sensor, and required the sensory data to be artificially clustered to produce spatial estimates at different spatial scales. This work investigates the utilization of multiple sensors within the multi-scale framework.

In this paper, we propose a new multi-scale mapping system that fuses place recognition hypotheses from different sensing modalities that naturally vary in spatial acuity in order to improve place recognition performance, akin to the multi-sensory, variable spatial resolution framework which has been discovered within mammalian brains.

3 Approach

In this section, we describe the three main parts of our approach; the creation of a multi-scale place cell representation, the multi-scale place recognition technique and our sensor fusion approach. We leverage biological inspiration from rats to improve place recognition for robotic platforms. The approach presented is an approximation of the multi-sensory place cell framework akin to the processes discovered within the mammalian hippocampus, enabling the encoding of places using multiple sensory modalities with a variable spatial resolution. The approach highlights the symbiotic relationship between both multiple sensory modalities and the multiple spatial resolutions within rodent navigational processes and demonstrates robust navigation within a robotics localization domain.

3.1 Map creation

Although some approaches posit a number of theories for place cell formation (Solstad et al. 2006; Spiers and Barry 2015; Barry and Burgess 2014), the actual mechanism of place cells creation and the method with which they become sensitive to specific spatial areas at specific spatial resolutions are currently unknown. In this work, we create the place cell architecture manually, creating “cells” which are spatially sensitive with varying resolutions.

Sensor data, from multiple sensors, are collected from an initial traverse of the environment (the “reference” traverse) enabling the storage of environmental information to create a map of the environment. Once the map is created, it can be used for localization on subsequent traverses of the environment. The sensor data are preprocessed in the same manner for both the reference and query traverses (see Sect. 4.3). Here, we present the place learning technique applied to each sensor.

For each individual sensor, we create cells with specific spatial resolutions by temporally clustering sensory data captured from the reference traverse. Traditionally, within the literature, place cell activations within an environment are represented as a Gaussian function (Muller et al. 1987; Keefe and Burgess 1996). In this model, the sum of these Gaussian functions is used to estimate the location of the agent within an environment; however, these techniques primarily seek to explain the correlation that exists between the location of rodents and the spatially specific activations of neurons within the hippocampus, rather than the cause of activation of these neurons within the world.

In our approach, we create place cells from stored sensory snapshots captured from the reference traverse. The reference traverse is divided into $N$ clusters, each with $M$ sensor samples per cluster. Our proposed technique encodes three core values within each place cell. We encode the spatial mean position, $\mu_{\text{RF}}(n)$, and the spatial standard deviation, $\sigma_{\text{RF}}(n)$, of each cluster representing the physical location within the environment each cluster represents, or the receptive field of each of the place neurons, as is common within the literature. However, current literature only attempts to model the activation behaviors measured by place cell neurons, neglecting to model the sensory inputs which cause these cells to become active. Utilizing only two parameters creates a disconnect between perceived sensory data and the location of the agent within the world; there exists no model of the stimuli required to activate the neuron; specifically, there is no method to perform data association between the incoming external sensor data and the place in the world. We extend the traditional representation of place cells and include an additional parameter, the average “appearance” or average sensory representation which encodes the requisite input sensory information to activate the place cell within an environment, with the place cell firing maximally when the agent is within an environment which is similar based on available sensory readings.

We represent the average appearance of each place cell as a radial basis function (RBF), using the average appearance parameter as a method for activating particular place cells based on incoming sensory data during the query traverse. The sensory appearance of each cluster is compressed and transformed by taking the mean for each stored feature within the cluster using

$$
\mu_k(n) = \frac{\sum_{m=1}^{M} R_k((n-1)M + m)}{M},
$$

where $n \in \{1, 2, \ldots, N\}$, $k \in \{1, 2, \ldots, K\}$, where $K$ is the length of the sensory feature vector, and $R_k(i)$ is the $k$th feature response of the reference snapshot $i$ along the...
Our approach presents a framework to combine sensors with variable place resolutions. A single sensor can be utilized at multiple spatial scales, or multiple sensors with multiple spatial scales can be used. Our approach combines place match hypotheses over a range of scales obtained from each sensor individually to obtain a global place match hypothesis reference traverse. Clusters are utilized to vary the spatial resolution of each sensor; larger clusters represent larger spatial areas, while smaller clusters represent smaller spatial areas, as shown in Fig. 4.

With this three parameter place cell, it is possible to recognize a location within the world using sensory data and estimate the location of the agent.

### 3.2 Place recognition

Once the place cell map has been created, it can be utilized for place recognition. The sensor data for the query traverse are preprocessed in the same manner as the reference traverses (see Sect. 4.3). Here, we present the place recognition technique applied to each sensor.

As the robot moves through the environment, each new query frame is compared to the place cell map described above. For each new query frame, the similarity of the current sensory information to the previously learnt place cells is determined by using a radial basis function (RBF) kernel as follows:

$$S(n, q) = \frac{1}{K} \sum_{k=1}^{K} e^{-\frac{|Q_k(q) - \mu_k(n)|^2}{2\sigma^2}},$$

where $S$ is the similarity score between reference cluster $n$ and query snapshot $q$, $K$ is the length of the feature vector and $\mu_k(n)$ is the mean cluster appearance for feature $k$ of cluster $n$. $Q_k(q)$ is the $k$th feature element of query sensor snapshot $q$. $\sigma$ is a free parameter which is set to 0.5 for all sensors and experiments. The RBF kernel ensures that all values are scaled between 0 and 1 with a strong match equal to 1.

### 3.3 Sensor fusion

Here, we present the technique to fuse the place cell activations generated from multiple sensors and multiple spatial scales. To achieve sensor fusion in this system, we represent the location of each sensor sample as a normal distribution, as is common within the place cell literature (Muller et al. 1987; Keefe and Burgess 1996), storing the mean, $\mu_{RF}(n)$, and standard deviation, $\sigma_{RF}(n)$, of each cluster or place neuron, encoding the spatial resolution of the sensor reading. Encoding each sensor sample as a normal distribution enables the fusion of sensor modalities regardless of the spatial or temporal resolutions and enables the generation of place recognition hypotheses at arbitrary interpolated locations:

$$E(r, q) = \sum_{n=1}^{N} \frac{1}{\sigma_{RF}(n)\sqrt{2\pi}} e^{-\frac{(r-\mu_{RF}(n))^2}{2\sigma_{RF}(n)^2}} S(n, q),$$

where $E(r, q)$ is the place cell activation for a given query location $q$ and location within the reference traverse $r$. Locations $r$ and $q$ are one-dimensional positions corresponding to locations along the captured reference path. These points may be grounded to 3D locations; however, that is not investigated within this work. $R$ and $Q$ represent the number of reference and query sensor samples. The standard deviation for each cluster is determined based on the sampling period of the sensor and the number of samples within the cluster, giving an indication of the area which the sensor snapshot represents.

To account for the variability of sampling rate, we also model the temporal activation of each of the place cells, enabling them to gradually activate and then decay. We model this temporal activation as a Gaussian centered around each new query frame:
\begin{equation}
T(r, q) = \sum_{j=1}^{Q} \frac{1}{\sigma_q(j) \sqrt{2\pi}} \exp\left(-\frac{(q-\mu_q(j))^2}{2\sigma_q(j)^2}\right) E(r, q). \tag{5}
\end{equation}

The mean for each query frame, \( \mu_q(j) \), is the time the frame was captured and the standard deviation for each frame, \( \sigma_q(j) \), is the sampling period of the sensor (WiFi = 1 Hz, Barometer = 1 Hz, Camera = 2 Hz). This enables the smooth representation of place cell activations over time for sensor data with different sampling rates.

Finally, the combined sensor matching score \( C(q, d) \) is generated using

\begin{equation}
C(r, q) = \sum_{k=1}^{W} T_k(r, q), \tag{6}
\end{equation}

where \( W \) is the number of sensors combined and \( T_k \) is the interpolated similarity score for sensor \( k \).

### 3.4 Selecting a place match

To determine if a query location matches a reference location, a search is performed for the place hypothesis with the largest combined sensor matching score:

\begin{equation}
b(q) = \arg\max_{r \in R}(C(r, q)), \tag{7}
\end{equation}

where \( b(q) \) is the best place hypothesis for query location \( q \) and \( R \) is the set of all locations in the reference traverse.

The current best matching sensory snapshot is determined to be a place match if the difference score is above a global matching threshold \( s_{\text{thresh}} \):

\begin{equation}
m = \begin{cases} 
1, & C(b(q), q) \geq s_{\text{thresh}} \\
0, & C(b(q), q) < s_{\text{thresh}}.
\end{cases} \tag{8}
\end{equation}

It is this threshold, \( s_{\text{thresh}} \), which determines if a particular location is a place match. \( s_{\text{thresh}} \) is swept over a range of values to generate the precision–recall curves presented in the results. Autonomous calibration of this threshold has been investigated in Jacobson et al. (2015a, b).

### 4 Experimental setup

In this section, we describe the hardware used for data acquisition and the testing environments used for evaluation of our proposed system.

#### 4.1 Testing environments and dataset acquisition

We evaluated our algorithm utilizing a dataset captured within a university campus, shown in Fig. 5. The dataset consists of three 1.5-km-long traverses of the campus environment during the day and night, including traversing two

---

*Fig. 5* Topological map (a) and an elevation map (b) for the campus environment with example images (i–vi) from the day and night traverses
levels of a building, descending and ascending 4 flights of stairs, and traversing outdoor environments ranging from natural foliage areas to building alleyways. Many areas of the testing environment underwent significant appearance change due to lighting changes and pedestrian traffic. The first traverse was captured in the morning and is considered the reference dataset or traverse. The second traverse was captured at night (referred to as the night dataset), and the third traverse was captured the next day (referred to as the day dataset). Each traverse was captured using standard laptop capable of capturing WiFi information, an external web-camera and a barometer using the Robot Operating System (ROS). The dataset is composed of 3 ROSbags, one for each traverse, and is available for download at https://wiki.qut.edu.au/display/cyphy/Datasets.

4.2 System evaluation

As with other place recognition studies, we used a precision–recall metric (Cummins and Newman 2011) and area under the curve (AUC) metric (Bonardi et al. 2017) for evaluating the place recognition performance of the system for the day and night test traverses. Precision is a measure of the number of correct place matches divided by the total number of place matches reported. Recall is measured as the ratio of correct place matches to all of the recallable place matches within a dataset. A precision–recall curve is a two-dimensional plot with the X-axis indicating recall and the Y-axis indicating precision. Measuring the maximum recall at 100% precision provides a measure of the percentage of the dataset that can be correctly recognized with no false-positive matches. Ideal performance is achieved when precision and recall are equal to 100% indicating all potential matches have been reported and all reported matches are correct. The AUC metric is the integral of the precision–recall curve and is another method to compare performance of different implementations, alleviating the potential brittleness of the maximum recall at 100% precision metric. Maximum performance of the AUC metric occurs at 100%, with the AUC metric giving an indication of precision over all recall levels.

For this work, true and false positives were labeled using a tolerance of 10 m, consistent with other research in this area. Ground truth maps for the campus datasets were calculated by manually labeling frame correspondences between each traverse. We also present a comparison to OpenFABMAP (Glover et al. 2012), comparing the performance of our system to a conventional single camera place recognition system which utilizes a single spatial scale. Finally, for scenarios where localization latency is not critical, we also include studies that apply the sequence-based place recognition system SeqSLAM (Milford and Wyeth 2012) to the place match hypotheses output by the multi-sensor, multi-scale place recognition system, replicating the temporal and spatial filtering evident in the entorhinal–hippocampal formation. The SeqSLAM algorithm utilizes a “sequence length” parameter which determines the amount of prior sensory information utilized in creating a place match, with a sequence length of \( N \) utilizing the past \( N \) sensory samples in generating a place match.

4.3 Preprocessing

The preprocessing techniques used here are to facilitate the integration of multiple sensor modalities and enable the utilization of a single place recognition framework. The preprocessing steps for each sensor are performed to each sensor sample from the “reference” and the “query” datasets. For each of our three sensors, the camera, WiFi and barometer sensors, preprocessing is performed to transform data into a single resultant format which can be utilized by the proposed system.

4.3.1 Camera preprocessing

The camera preprocessing step is a standard feature extraction; we extract histogram of oriented gradients (HOG) features for each image (Dalal and Triggs 2005). The HOG feature is a common feature type used for place recognition (Lategahn et al. 2013). HOG features are extracted for each database and query image; with the feature set for each query image being compared to database features to produce a place hypothesis, as described in Sect. 2.2. We use full-size grayscale images at 640 × 480 resolution and use a block size of 64 for extracting the HOG features.

4.3.2 WiFi preprocessing

WiFi place recognition leverages available WiFi access points and the corresponding received signal strength (RSS) from each device to create a unique “fingerprint” for each location in an environment. The preprocessing for the WiFi sensing modality involves creating a database by collecting a list of all detected WiFi access points captured during the database traverse of the environment. The signal strength for each detected WiFi point is recorded for each place in the database. During the query traverse, only WiFi access points which match the access points found during the database traverse are recorded, compared and used for place recognition.

4.3.3 Barometer preprocessing

Place recognition using a barometer has not yet been proposed in the robotic SLAM literature. The implementation of barometric sensor within this work leverages the relative change in pressure, relative to a known location, which is constant across an environment regardless of the current
absolute pressure. It is effectively giving an estimate of elevation (Li et al. 2013), an estimate which is reliable over multiple traverses of an environment, both indoors and outdoors. The sensor is not commonly used for localization (which makes it appealing for demonstrating the utility of the proposed approach) due to the fact that it can only represent an elevation in an environment and hence typically produces aliased results, especially when operating in environments where there is a lack of altitude change. This is not a problem within the proposed framework due to the utilization of multiple spatial scales by our multi-sensor fusion approach. Place recognition using a barometer is enabled by using the relative pressure between a known location and the pressure at the agent’s current location:

\[ P_t = P_1 - P(j), \]  

(9)

where \( P_1 \) and \( P(j) \) are the pressure at the known location and the current pressure at the agent’s location, respectively. The relative pressure is used to account for the fact that barometric pressure can change, depending on the time of day and weather conditions. Requiring pressure data from a known location does not assume that the agent knows where it is at the beginning of the localization task, nor does it assume the agent has to go to this location. It is acceptable for this measurement to be collected by a third party and relayed to the navigating agent.

5 Results

Here, we present the results for the proposed multi-scale sensor fusion technique. We present results comparing the performance of the individual sensor place recognition techniques and the proposed multi-scale sensor fusion technique. For both day and night datasets, we present precision–recall and AUC graphs for individual sensors and the fused multi-scale system. Furthermore, we include a comparison to FABMAP, investigating the benefits of multi-scale approaches when compared to a single-scale system. We also analyze the characteristics and spatial acuity of the different sensing modalities and different spatial resolutions.

5.1 Multi-scale utility

We demonstrate the impact of utilizing multiple scales and multiple sensors by comparing the percentage recall at 100% precision for the day (Fig. 6) and night (Fig. 7) datasets. It can be seen that for any static sequence length, combining two sensors for all cluster sizes improves what is possible with a single sensor; this is evident both for the day dataset (comparing the results in Fig. 6a–c and d–f) and the night dataset (comparing the results in Fig. 7a–c and d–f). Furthermore, fusing three sensors does provide better results than what is possible with any individual sensor or two-sensor combination (Figs. 6g, 7g). Figures 6g and 7 also illustrate that, for a given sequence length, increasing the cluster size used can improve performance; however, there is a maximum to the performance increase attained. If a cluster size is too large, performance will start to degrade again, this is attributed to both the clusters representing a spatial area which is too large (creating place matches which may be correct, but are centered a distance away from ground truth) and clusters attempting to represent too much sensory data. It is also worth noting that for a given cluster size, the utilization and fusion of two or three sensors increases the recall at 100% precision and enables the utilization of shorter sequence lengths (while achieving similar or superior localization performance over a single sensor) reducing localization delay and increasing the feasible operating zone.

Precision–recall graphs are presented in Figs. 8 and 9. These graphs were generated using a sequence length of 20. The sequence length of 20 was selected from analysis of Figs. 6 and 7, a sequence length which in all plots did not produce a perfect result or 100% recall at 100% precision. The selection of a sequence length which does not achieve perfect results enables the analysis of the improvement in performance without experiencing any “clipping” of results. Furthermore, it is also worth noting that at this sequence length of 20, the performance for all individual sensors was less than 4% recall at 100% precision, a result which is unusable in a real-world deployment. However, at this sequence length, it can be shown that sensors which individually demonstrate poor performance can be fused to achieve superior results.

Figure 8 presents the precision–recall graphs for the daytime dataset. It can be seen that the integration of multiple sensors improves what is possible with a single sensor, comparing Fig. 8a, b and c, d and e, f and g, h; all fused results produce a result better than what is possible with one sensor. It can also be seen that combining multiple scales and multiple sensors in Fig. 8h produces better results than individual sensors. The best overall performance is attained in Fig. 8d using the camera, WiFi and barometer sensors with a cluster size of 5, achieving 26.05% recall at 100% precision. The second best result is attained by fusing camera, WiFi and barometer sensors with a cluster size of 5 and 10, achieving 24.16% recall at 100% precision. These results are an improvement over the best single-sensor result of 3.14% recall at 100% precision achieved with the WiFi sensor with a cluster size of 10.

The night-time dataset is very challenging for the camera modality due to the visual information completely changing, while the barometer and WiFi sensors remain relatively constant when compared to the day-time dataset. However,
Fig. 6 Maximum recall at 100% precision for the day dataset (yellow is higher recall). a Camera, b barometer, c WiFi, d camera and barometer, e camera and WiFi, f barometer and WiFi and g camera, barometer and WiFi combinations were evaluated. Combining all the sensors across all scales improves the feasible operating space by enabling shorter sequence lengths at all matching scales (color figure online).

Fig. 9a, b and c, d and e, f. The combination of both multiple scales and multiple sensors produces the best overall performance with increasing recall at 100% precision from...
Fig. 7 Maximum recall at 100% precision for the night dataset (yellow is higher recall). a Camera, b barometer, c WiFi, d camera and barometer, e camera and WiFi, f barometer and WiFi and g camera, barometer and WiFi combinations were evaluated. Although absolute performance is lower than for the day-time dataset, once again combining all sensing modalities across all scales results in an improved feasible operating zone by enabling shorter sequence lengths at all matching scales (color figure online).

2.02% with the WiFi sensor modality and a cluster size of 10 in Fig. 9e to 24.09% for the multi-scale, multi-sensor result in Fig. 9h.

Figure 10 shows the AUC graphs for the day and night datasets. The results illustrate, for both day (Fig. 10a) and night (Fig. 10b) datasets, that there is an improvement
Fig. 8 Precision-recall curves for the day-time dataset using SeqSLAM with a sequence length 20. Comparing single-sensor results to multi-sensor results, the integration of multiple sensors universally improves performance at all cluster sizes. The best performance attained is attained in (d) using the camera, WiFi and barometer sensors with a cluster size of 5, achieving 26.05% recall at 100% precision. The second best result is attained in (b) by fusing camera, WiFi and barometer sensors with a cluster size of 5 and 10, achieving 24.16% recall at 100% precision.
Fig. 9  Precision–recall curves for the night-time dataset using SeqS-LAM with a sequence length 20. Under challenging conditions at night, camera performance is much degraded, while WiFi and the barometer stay relatively constant. Similarly to the day-time dataset, the utilization of multiple sensors improves localization performance. The best performance of 24.09% recall at 100% precision is attained with multiple sensors and multiple scales (2,5,10) in h.
The results clearly show, for both day and night datasets, the improvement of multi-sensor configurations over single-sensor configurations. The results also illustrate that the multi-scale, multi-sensor system outperformed the results from multi-sensor or multi-scale configurations in isolation.

It can be seen in the AUC graph that the performance improvement from utilizing multiple sensors and multiple scales is achieved across all multi-scale configurations across both the day and night datasets. The single-scale configuration for camera, barometer and WiFi with a cluster size of 5 approaches or exceeds the multi-scale, multi-sensor results for both day and night datasets. It should be noted that by leveraging multiple sensors and multiple scales—even scales which do not perform optimally in isolation—once combined achieve superior results to any of the individual sensors, indicating that while a smaller set of scales can be used, using multiple scales can ensure robust performance without performing analysis at multiple individual scales to determine optimal scale parameters.

5.2 FABMAP comparison

We include a comparison to FABMAP in Fig. 11 with using both the day-time and night-time datasets. FABMAP was
Recall(%)

0 20 40 60 80 100

Precision(%)

Precision-Recall Graph

Day
Night

Fig. 11 Precision–recall curves using FABMAP for the day and night datasets, included for completeness. Understandably, performance is hampered by the poor condition invariance of the feature detectors trained using the reference traverse in an attempt to favor the performance of the FABMAP system. It can be seen that for both the day-time and night-time the performance of the FABMAP system does not achieve 100% precision. Furthermore, our proposed system outperforms FABMAP for all multi-sensor fusion combinations during the day-time dataset except for the multi-sensor results using a cluster size of 2, Fig. 8f, using WiFi and camera sensor modalities where the precision does not reach 100%. Our approach performs similarly at night when compared to FABMAP, achieving superior results during the night traverse with all sensor fusion combinations achieving 100% precision.

5.3 Spatial acuity analysis

Figure 12 shows the relative percentage contribution of each sensor toward correct place matches at each place scale, normalized so that each sensor contributes a total of 100 percent across all scales, for day and night tests and with regard to instantaneous place recognition performance as well as sequence-based performance.

Remarkably, all three sensors display the same spatial acuity characteristics regardless of day or night conditions, and regardless of whether single snapshot-based or sequence-based matching is performed. When comparing the components of each correct place match, the camera sensor is consistently more informative at performing place recognition at spatially specific scales, followed by the barometer and then WiFi.

| Sensor Contributions — a day (sequence length 1), b night (sequence length 1), c day (sequence length 20) and d night (sequence length 20). These results illustrate that when comparing the components of each correct place match, the camera sensor is consistently more informative at performing place recognition at spatially specific scales, followed by the barometer and then WiFi. The spatial information provided by sensors is consistent across day and night datasets and with varying sequence lengths |
5.4 Weighted spatial acuity study

Weighted spatial acuity provides a method for variably weighting sensor scales to further improve the place recognition performance of the multi-scale framework. We generate scale weightings using the spatial acuity of a spatial scale for a particular environment and its ability to generate correct place matches according to the spatial acuity analysis. These weightings are environmentally specific; future work will investigate the usage of online calibration or dynamic sensory weighting techniques (Jacobson et al. 2015a, b). The weightings are applied to each scale during the sensor fusion process in Eq. 6 by multiplying each spatial scale by the spatial acuity value for each scale.

The weighted multi-sensor, multi-scale fusion, shown in Fig. 13, produces a result which further improves performance over what is possible with individual sensor modalities, achieving 37.45% recall at 100% precision in the day-time dataset. The night-time dataset also shows an improvement over what is possible with individual sensors, achieving 26.72% recall at 100% precision. The results from the AUC graph, shown in Fig. 14, show that the weighted multi-sensor, multi-scale system outperforms what is possible with an individual sensor or naive multi-scale, multi-sensor combination. For the night dataset, the AUC performance outperforms what is possible with individual sensors; however, it does not outperform a naive multi-sensor, multi-scale system. The weighted multi-scale system does not outperform a naive multi-sensor combination due to erroneous place matches from the camera sensor receiving a stronger weighting than other sensors; future work will investigate online sensor weighting to remove erroneous place matches (Jacobson et al. 2015a, b).

6 Discussion and future work

This paper presented a biologically inspired multi-scale, multi-sensor place recognition system that incorporates the varying spatial localization estimates provided by different sensing modalities to improve overall place recognition performance. We focused on developing a model of place cells discovered within the rodent hippocampus, incorporating the multi-scale, multi-sensory nature of the place cells into our robotic localization model. The models proposed within the literature primarily aim to model the correlation between spatially specific neuron firing and the location of a rat within an environment. Within this work, we expanded traditional place cell models to enable utilization within a robotics domain, including a parameter which encodes the average sensory representation of a place. The inclusion of the average sensory representation of a place enables our proposed model to both represent a “place” within the world, but also enables the spatially specific firing of place cells according to current sensory information.

In experiments across a university campus, both outdoors and indoors over multiple building floors, place recognition performance was universally improved when individual place recognition estimates from a camera, WiFi receiver and barometer were combined over multiple spatial scales. Notably, while combining sensors at a single scale generally improved place recognition performance over using individual sensors, a significant further improvement was obtained by combining these same sensors using multiple different recognition scales.
The multi-scale framework enabled us to analyze how each sensor contributed to place recognition at different scales, revealing that different sensors exhibited significant and consistent variations in the scale at which they were most spatially informative. This insight explained why using a weighted multiple scale framework further improved the multi-sensor place recognition performance, by ensuring all sensors were captured at their optimal scale range. By weighting sensors by their utility for localization at each scale, we demonstrated how this sensor characterization can result in a further significant improvement in place recognition performance.

In biology, it is clear that animals, such as rodents, are known to have spatially responsive cells and can use multiple sensing modalities. There has been significant work attempting to identify the contributions of various sensing modalities to spatial encoding in the brain, especially in specific circumstances such as when the lights are turned out (Hargreaves et al. 2005; Rossier et al. 2000; Sheppard et al. 2013; Maaswinke and Whishaw 1999). The research presented here suggests that there may be other critical differentiators for how and to what extent sensing modalities are used. In this work, we proposed an adapted model of place cells and presented results which illustrate improvements within a robotics domain and may present insights into what may be occurring within place cells to imbue them with the spatial selectivity observed by biological researchers.

In this work, we analyzed the utility of different sensing modalities at different scales; such an analysis on rodent or other mammalian brain recordings might yield insights into which sensing modalities are used by an animal under what circumstances. The analysis performed here also showed that place recognition is more readily performed at different scales depending on the context and sensing modality—in rodent recordings, we might expect a remapping event to initiate at one particular spatial scale and then propagate to the other spatial scales, although observing such a phenomenon might be difficult using current techniques. Finally, despite the difficulty in performing experiments, there are potential new insights to be gained from recording from animals moving over large distances; preliminary work recording from bats over long distances may reveal new insights into how multiple sensing modalities inform spatial mapping over multiple scales (Geva-Sagiv et al. 2015).

One of the most obvious areas for future investigation is to comprehensively investigate the utility of all commonly available commodity sensors (such as those found on a modern smartphone) for place recognition. Sensing modalities including sound and light sensors may be nearly useless for place recognition when used in isolation, but may provide powerful contextual information when combined with multiple other sensors and at multiple scales. This approach holds promise because the results in this paper show that the barometer, while being a poor localization sensor in isolation, is able to improve place recognition performance when combined with other sensing modalities.

The range of scales over which a multi-scale approach is viable is also unknown, mirroring the uncertainty in neuroscience as to how large a scale can be encoded by the multi-scale neuronal map found in the brain. By scaling up experiments to occur over environments ranging in size from single rooms to cities or larger, we will be able to evaluate whether the multi-scale approach continues to scale indefinitely. Evidence from migratory animals such as terns suggests that up to four or more different scale mapping strategies are used in traversing the globe, suggesting global scalability is achievable (Frost and Mouritsen 2006; Geva-Sagiv et al. 2015).

Finally, a multi-scale approach offers much potential for computational optimization. Reliable but spatially non-specific sensors such as WiFi may enable computation of place match hypotheses for other more spatially specific sensors such as cameras to only be performed in promising match areas, rather than globally throughout the entire environment. WiFi signatures are theoretically unique because of the uniqueness of MAC addresses, meaning it should be impossible to get an aliased match outside of the coverage range of a particular network except in the case of multi-path or environment modification. In future work, building on this foundation, we hope to enable robust navigation for low-cost robot platforms equipped with cheap commodity sensors through the combination of improved place recognition performance and reduced computation provided by this type of multi-scale, multi-sensor approach.

References

Andreasson H, Duckett T, Lilienthal A (2008) A minimalistic approach to appearance-based visual SLAM. IEEE Trans Robot 24(5):1–11
Arras KO, Tomatis N, Jensen BT, Siegwart R (2001) Multisensor on-the-fly localization:: precision and reliability for applications. Robot Auton Syst 34(23):131–143. https://doi.org/10.1016/S0921-8890(00)00117-2. ISSN 0921-8890. European Workshop on Advanced Mobile Robots
Barry C, Burgess N (2014) Neural mechanisms of self-location. Curr Biol 24(8):R330–R339
Bay H, Tuytelaars T, Van Gool L (2006) SURF: speeded up robust features. In: Computer vision ECCV 2006
Berkvens R, Jacobson A, Milford M, Peremans H, Weyn M (2014) Biologically inspired SLAM using Wi-Fi. In: IEEE intelligent robots and systems
Bonardi F, Ainouz S, Boutteau R, Dupuis Y, Savatier X, Vasseur P (2017) PHROG: a multimodal feature for place recognition. Sensors 17(5):1167
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, vol 2. IEEE, Taipei, pp 1899–1906. ISBN 0780377362
