Gist+RatSLAM: An Incremental Bio-inspired Place Recognition Front-End for RatSLAM

S. M. Ali Musa Kazmi
Department of Electrical Engineering
University of Paderborn
Paderborn, Germany
kazmi@get.upb.de

Bärbel Mertsching
Department of Electrical Engineering
University of Paderborn
Paderborn, Germany
mertsching@get.upb.de

ABSTRACT

There exists ample research exploiting cognitive processes for robot localization and mapping, for instance RatSLAM [10]. In this regard, tasks such as visual perception and recognition, which are primarily governed by visual and perirhinal cortices, receive a little attention. To bridge this gap, we present a novel bio-inspired place recognition front-end for the RatSLAM system. Our algorithm uses Gist features to obtain the perceptual structure of the scenes and employs a modified growing self-organizing map (GSOM) to model the behavior of the cells found in perirhinal cortex, called recency and familiarity neurons [6]. This enables an online learning and recognition of the places without acquiring apriori knowledge of the environment. The experiments carried out on the standard St. Lucia dataset demonstrate that on average our approach achieves almost 10% improvement (in F1-Score); it is able to correctly flag the visited and unvisited places even for noisy and blurred visual inputs. The results show that the proposed method reaches fast convergence and utilizes a smaller number of cells (consumes less physical memory) to represent the traversed path compared to the RatSLAM approach.

Keywords

Gist, bio-inspired mapping, scene recognition, global image feature, self-organizing neural network, competitive learning

1. INTRODUCTION

To produce a meaningful representation of an environment, a mobile robot should be able to distinguish visited and unvisited places. Classical methods aim to tackle this problem by keeping a short history of landmarks' positions and thus minimize the localization errors based on a multi-hypothesis approach (i.e. particle filter) [27]. In the same vein, pose graph optimization algorithms [28, 7] solve a non-linear least squares problem using conjugate gradient descent to approximate an optimal representation of the environment. Even though these approaches offer fairly good Cartesian maps, such as a trajectory of the estimated robot motion [29] or occupancy grid maps [12], they lack the notion of place recognition.

Recent developments in computer vision have led the researchers to solve the robotic mapping problem in the domain of scene recognition; examples include appearance-based mapping [13, 8, 3]. The mainstream of these algorithms uses a bag-of-words (BoW) model, introduced by Sivic et al. [23], to learn a visual vocabulary of SIFT or Gist features. Others impose 3D reconstruction constraints between a pair of images to group similar places into a single cluster, see for instance [31]. Cummins and Newman presented a probabilistic formulation to use visual words, computed from quantized SURF descriptors, for fast appearance-based mapping of 1000 km long trajectory (FAB-Map) [3]. The vocabulary is built offline with a Chow Liu tree, and it is later used to approximate the likelihood of being at a particular location based on the probability of which words are observed in the image. A BoW model has also been adopted to learn Gist features for different similarity measures [13]. The authors suggested a randomized k-d tree based approach as a feasible choice, which is trained on projected Gist features obtained from PCA. The complexity of the algorithm is $O(logN)$, where N is the number of images in the database; training multiple k-d trees did not show any improvements. A prior stage of learning the visual words makes these methods less suitable for unknown environments.

Kawewong et al. introduced an incremental approach to learn the BoW model for position invariant robust features (PIRF) [8], derived from SIFT features appearing in consecutive frames over a sliding window. Their method showed a better performance than FAB-Map and is robust for dynamic environments. Nevertheless, it is dependent on the window size and for an appropriate window size the average computation time is 2 to 3 seconds. In an attempt to do online place recognition, Suenderhauf and Protzel used BRIEF descriptors around the center of downsampled image, which they called BRIEF-Gist, and employed it as a front-end for their pose graph optimization algorithm [25]. Unlike FAB-Map, it suffered weak data associations when a vehicle traversed the same place from a different direction.

In this paper, we exploit the concept of "Gist of the scene" formed in early vision and model a new scene learning mechanism for RatSLAM [11]. Unlike existing work, our method performs bio-inspired scene learning which enables our algorithm to map unknown environments. Further discussion in the paper proceeds as follows: Section 2 gives a brief overview of the related work. Section 3 describes the Gist

.

BICT 2015, December 03-05, New York City, United States
Copyright © 2016 ICST
DOI 10.4108/eai.3-12-2015.2262532

descriptor to obtain the structural formation of the scenes. This information is later fed to the modified growing self-organizing map (GSOM), a type of neural network, to organize the places based on their perceptual distances. A comprehensive evaluation of the proposed approach on the standard St. Lucia dataset is presented in Section 4, including comparison with the state-of-the-art RatSLAM algorithm. The results show that the proposed approach performs better than existing front-end module (data association) of the RatSLAM system and imposes less consumption of physical memory. Finally, we conclude the discussion in Section 5 with the possible directions of future work.

2. RELATED WORK

In an early vision, a scene is represented by low-level features (e.g. color opponency, oriented edge, etc.) at different spatial scales. This information is held temporarily in *iconic memory* and is only transferred to long-term memory if the control processes (i.e. attention, rehearsal) decide to keep that information [6]. As this memory is highly volatile, it retains information only for a fraction of a second. Current studies on visual perception report that humans interpret the meaning of a scene within 200 ms of its presentation. The amount of information extracted during this period is referred to as "Gist" [18], provided the eye fixations or exposures to a new scene are separated by a gap of a few milliseconds. This indicates that a precise classification of the constituent objects of a scene is not needed in early vision [16]. In essence, such a holistic view of the scene could be of significant importance, for example in robotic mapping, and would augment the formation of spatial memory for long-term cognitive mapping.

In this respect, Siagian and Itti developed a model to compute Gist features and saliency regions in parallel from saliency maps, which are then fed to a trained back-propagation neural network and SIFT recognition module, respectively, for place recognition [22]. At the back-end, a variant of the Monte Carlo method is implemented to estimate the most likely position of a robot. The experiments are done on small scales and the recognition process took almost 3 seconds at a machine with a 16-core 2.6 GHz processor. Tapus and Siegwart combined features from different modalities (i.e. a laser and an optical camera) to form the fingerprint of a place for topological mapping based on a POMDP (Partially Observable Markov Decision Processes) framework [26].

Milford and Wyeth demonstrated the RatSLAM algorithm on a 66 km long suburb of St. Lucia [11]. Their work draws upon the models of Arleo and Gerstner [1] with major modifications in the model for place and head direction cells, the type of cells found in a rat's brain which fire maximally when animal is at a particular place or facing a specific direction [15]. The system learns associations between scenes and pose cells – a network of place and head direction cells; path integration in the pose cells network is driven by translational and angular velocities while dead-reckoning errors are re-calibrated via visual input. The performance of the system relies on the parameter settings and the size of a pose cells network. Glover et al. integrated FAB-Map with the RatSLAM to improve recalls and reduce the overhead of configuring parameters [5]. Their results showed that offline vocabulary of SURF features is sensitive to illumination and does not remove the parameters' dependency. The work of Milford and Wyeth has also been extended to implement a security system that uses a dendritic cell algorithm (DCA) for anomaly detection in the environment [14].

Chen et al. modeled the concept of multi-scale spatial maps discovered in rodent's brain [2]. They trained arrays of SVMs (Support Vector Machines) on Gist features for overlapping segments along the path at different spatial scales. For a query image, hypotheses from arrays of SVMs are combined to perform place recognition. Their experiments show that there is no rule of thumb that multi-scale place recognition would always outperform single scale recognition. Rather, there exist certain cases where single scale place recognition works better. This work could although suggest a good method to recognize places at multiple scales, but it has not been demonstrated to perform mapping. Additionally, an offline training on a dataset makes this approach biased to the learned environment and thus it does not seem to contribute significantly to relax the parameter tuning.

Some researchers have compared their approaches with RatSLAM. For example, Suenderhauf and Protzel formulated dynamics of pose cells in terms of a Bayes filter, which they named Causal Update Filter (CUF) and used a *TORO* pose graph algorithm for experience mapping [24]; however, no significant improvements are achieved compared to RatSLAM. Rebai et al. used a Fuzzy ART network to capture the properties of spatial view cells in primates [20]. The network is trained incrementally on the quantized local histograms of hue and saturation. It has been shown that their method outperforms RatSLAM regarding loop closure detection, but this is not demonstrated for mapping.

3. GIST BASED PLACE RECOGNITION

The human vision system characterizes places using different spatial frequencies, at several scales and orientations, without an explicit need of grouping the objects [18]. This global information of a scene could serve as a basis to construct its human-like coarse representation. This alone does not suffice the need of human way of recognizing places. Therefore, we attempt to model the behavior of recency and familiarity neurons using a growing self-organizing map (GSOM) to learn and recognize places.

3.1 Computing Gist of a Scene

The Gist of a scene represents the structural properties composing it. This information can be obtained from perceptual attributes, such as the degree of naturalness, openness, verticalness, etc. These features are shared among places and thus aid to achieve continuous categorization of the scenes such that the places having similar attributes lie close to one another on the perceptual axis, as shown by Torralba and Oliva [17]. They computed these features by sampling the energy spectrum $A(f_x, f_y)^2$ of an image $I(x, y)$ using a set of Gaussian functions $G_i(f_x, f_y)$ at different orientations and scales:

$$g_i = \int\int A(f_x, f_y)^2 \, G_i(f_x, f_y) \, df_x \, df_y \tag{1}$$

The Gaussian functions G_i model the Gabor filters like responses at different orientations and scales of spatial frequencies; they are obtained as follows:

$$G(f_x, f_y) = e^{-f_y^2/\sigma_y^2} \left(e^{-(f_x - f_0)^2/\sigma_x^2} + e^{(f_x + f_0)^2/\sigma_x^2} \right) \tag{2}$$

where f_0 specifies the center of the response function. The

parameters σ_x and σ_y control the scale of Gaussian in horizontal and vertical directions, respectively, for specified spatial frequencies f_x and f_y. Thus, the sampled energy spectrum of an image is represented by a vector $\mathbf{g} = \{g_i\}_{i=1:L}$, where L is the dimensionality of the feature vector. This makes Gist features suitable to build the human-like perceptual representation of the scenes.

3.2 Growing Self-Organizing Map (GSOM)

The cells found in the visual cortex and the associated areas of hippocampus exhibit a competitive response to represent an input pattern [9]. The type of neural network modeling such a map of neural activity in cortical regions of the brain is called as *Kohnen's Self-Organizing Map (SOM)*. In practical, realizing such a neural network to map environments with unknown size is computationally expensive. As a result, we selected a *Growing SOM* [21] to adapt to the dynamic size of the environments. The number of neurons in a GSOM tends to vary over time to adapt to the topology of the input space according to its size, as shown in Fig. 1. A neuron $i \in \{1, 2, 3, ..., m\}$ that closely resembles the observation $\mathbf{x}^{(k)} \in \mathcal{R}^n$ is deemed to be a winning neuron or the best matching unit and thus contributes strongly to represent it. The winning cell having the minimum distance from the kth input is determined as follows:

$$c^{(k)} = \underset{i}{\operatorname{argmin}} ||\mathbf{x}^{(k)} - \mathbf{w}_i|| \tag{3}$$

where $c^{(k)}$ is an index to the winner neuron mapping the kth input and $\mathbf{w}_i$ is the current weight of neuron i. In practice, one usually finds a set of best matching units for some $\mathbf{x}^{(k)}$ during the initial phase of learning. The arrival of further input leads to the convergence[1] of activity to a single winner neuron [4] where a group of neurons in the neighborhood $N_{c^{(k)}}$ of the winning neuron forms a receptive field. This requires adaptation of the weights for the winning neurons (along with the neighboring cells) given as:

$$\mathbf{w}_i^{t+1} = \begin{cases} \mathbf{w}_i^t + \alpha\,(\mathbf{x}^{(k)} - \mathbf{w}_i^t) & i \in \{c^{(k)}, N_{c^{(k)}}\} \\ \mathbf{w}_i^t & \text{otherwise} \end{cases} \tag{4}$$

where α is the learning rate. It can be constant, exponentially decreasing or inversely proportional to time t, depending upon the problem at-hand. The other considerations, such as an initial neighborhood size and initialization of weights are discussed in the next section.

3.3 Modeling the Front-End for RatSLAM

One of the neat characteristics of GSOM is that it does not require assumptions about the distribution of the feature space and the size of the network. These properties enable us to perform incremental scene recognition because we intend to learn from scratch in an unknown environment as soon as the stream of data is available, which in our case are image sequences, as shown in Fig. 2.

[1]The convergence of the activity to a winner neuron is inherent to network dynamics. However, it is subject to number of epochs (time), stationarity of the input, and convergence criteria. For general details the reader is referred to [4, 9]. The problem specific explanation is given in Section 3.3.3.

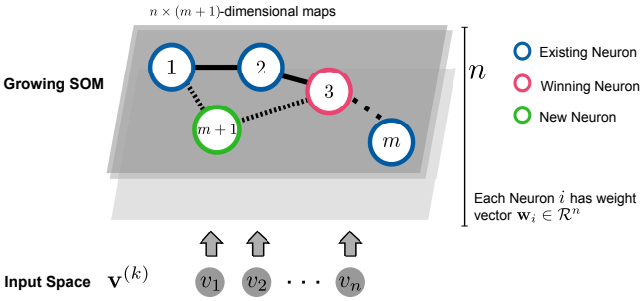

Figure 1: An example framework of Growing SOM illustrating the concept of the winner neuron and creation of new nodes in the network. The layers in GSOM, shaded gray planes stacked on top of each other, are maps of n-dimensional feature space.

3.3.1 Creation of GSOM Network

The creation of a GSOM network requires a) the decision of the initial size for the network and b) the way weights are initially assigned to start-up nodes. Usually, one starts with 3 to 4 initial nodes while weights of the nodes are randomly assigned or picked from the feature space. We start with only one node because in the start the only information about the environment one can have is the starting point. Moreover, the weight vector $\mathbf{w}_1$ of this node is set to the normalized Gist descriptor obtained from first image using (1). This is a justifiable consideration for the scenarios in which data arrives in streams as at this instant of the time nothing more can be known about an environment to be explored.

3.3.2 Distance Measure to Learn Input

When the kth input $\mathbf{v}^{(k)}$ is presented to the network, its distance $d_i^{(k)}$ is computed from each neuron i in a set of existing neurons $\mathcal{M}$. There exist several measures, such as Manhattan distance, Euclidean distance, Radial Basis, and others [4]. The selection of the distance measure is specific to the application and has a strong impact on the competition induced in the network. Here, the input space is composed of Gist features and they have been demonstrated to work fairly good for the sum of squared distances [17]. Therefore, the distance metric is given by:

$$d_i^{(k)} = \sum_{j=1}^{L} (v_j^{(k)} - w_{ij})^2 \tag{5}$$

where $d_i^{(k)}$ is the distance of ith neuron from the presented feature $\mathbf{v}^{(k)}$ and $L = 512$ is the dimensionality of the Gist descriptor.

3.3.3 Determine Best Matching Unit

A best matching neuron $c^{(k)}$ is the one whose distance from the presented feature vector $\mathbf{v}^{(k)}$ is minimum, which is obtained by minimizing (5) for all neurons i in a set $\mathcal{M}$ of the existing neurons. Since the outdoor environments are highly ambiguous in nature, it is likely to obtain almost a similar descriptor for two different physical locations. Consequently, it could lead to false positive recalls in the network. In order to reduce the likelihood of such an aliasing, we extend the distance measure to take into account the distances of neurons in the neighborhood of the winner cell. Therefore, we define the following objective function to determine the

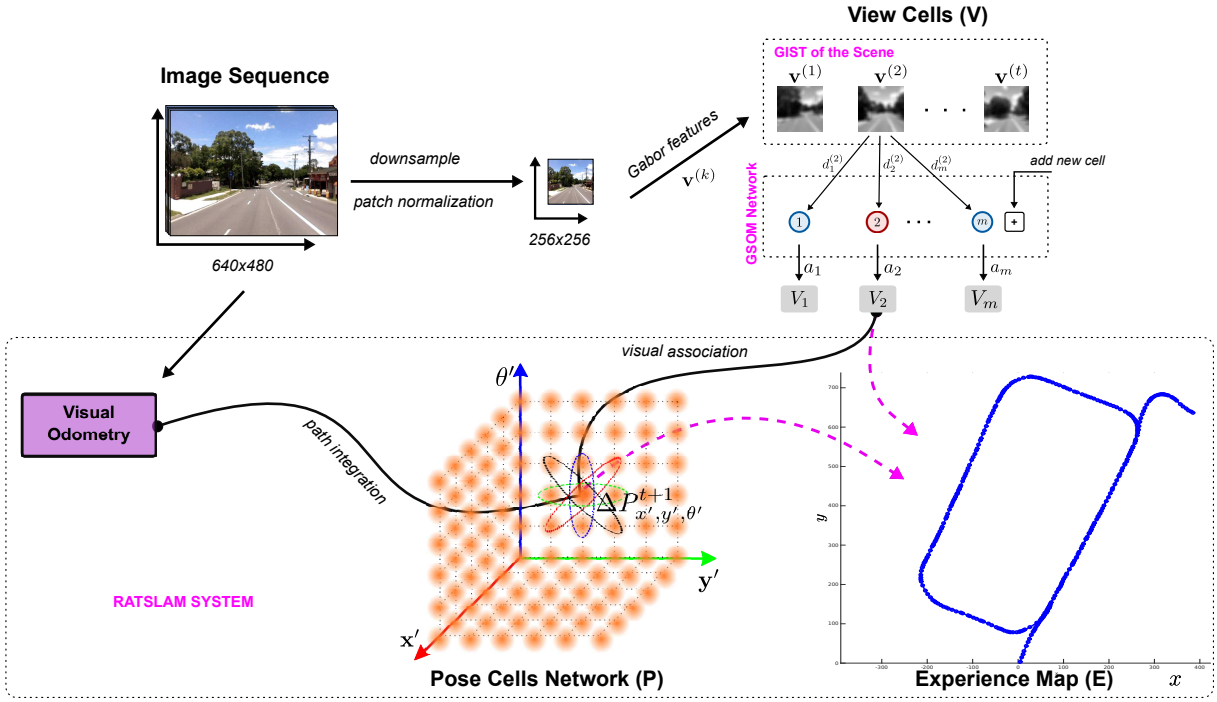

Figure 2: Gist+RatSLAM: GSOM based data association front-end for the RatSLAM algorithm. For each image in a sequence, its Gist features are computed and presented to the network. A neuron familiar to this input would show high activation (as indicated by a red circle in the GSOM network) being at minimum distance $d_2^{(2)}$. For a novel input a new neuron is created to learn the pattern.

best matching unit:

$$d_i^{(k)} = \min_i \left(d_i^{(k)} + \frac{1}{N} \sum_{\forall p \in N_i} ||\mathbf{v}^{(k)} - \mathbf{w}_p||^2 \right) \quad (6)$$

The summation on the right side of the equation acts like a penalizing term. It incurs the cost of making a wrong decision for associating different places; where N is the neighborhood size and $\mathbf{w}_p$ is weight of the neuron p in the neighborhood of the ith neuron. Usually, the size of the neighborhood is allowed to shrink with time, but it is not desirable for Gist features (as explained in the next section) so that only a particular neuron could be tuned to fire for specific perceptual attributes. It should be noted that finding an optimal best matching unit is not always guaranteed, because natural environments share a high-level of similarity. The time complexity of the proposed method to search for the best matching unit is quadratic in m; where m is the number of neurons in the network. Our algorithm is efficient compared to the RatSLAM algorithm in the sense that less number of experiences are created and therefore the search space has been reduced to find the winner neuron.

3.3.4 Adapting Weights of Neurons

Given the index of a winner neuron $c^{(k)}$, the next step is to adapt the weights $\mathbf{w}_{c^{(k)}}$ of the winning cell including its neighbors based on the learning rate α, see (4). Here, the GSOM algorithm benefits from the ability of Gist descriptors to segregate scenes as a contiguous organization along the perceptual axis. This suggests that during exploration places would appear consecutive to each other on the axis defined by the Gist features. On that account, we modified (4) to update the weights of the winner neuron only:

$$\mathbf{w}_i^{t+1} = \begin{cases} \mathbf{w}_i^t + \alpha\,(\mathbf{v}^{(k)} - \mathbf{w}_i^t) & i = c^{(k)} \\ \mathbf{w}_i^t & \text{otherwise} \end{cases} \quad (7)$$

The learning rate α of a neuron often decreases over time, whereas we opt to reduce the learning rate as a function of new nodes created in the neighborhood N_i of the neuron, it is defined as follows:

$$a^{t+1} = a_0\, exp\left(-\sum N_i/\rho\right) \quad (8)$$

$a_0 = 0.1$ is the initial learning rate and ρ is the allowed number of new nodes that can be created near a neuron i. The weights should satisfy the constraint $||\mathbf{w}_i|| = 1$.

3.3.5 Creating a New Neuron

The creation of a new cell needs to address several aspects based on the feature space, such as the position where a node should be created and the decision that the presented input is novel (i.e., no neuron in existing set $\mathcal{M}$ can represent it). In scenarios like exploration, the places are encountered as an ordered sequence. As discussed previously, in Sections 3.3.3 and 3.3.4, the Gist features allow a continuous categorization of the scenes along the perceptual axis. This implies that the GSOM network should grow in a sequential manner to learn new places. Hence, a new neuron is created next to the closest neuron (a neuron which necessarily satisfies (6)). Otherwise, for two neurons found to be almost equally closer to a presented Gist feature, we create a new neuron that is equidistant from them in a space defined by the Gist features.

3.3.6 *Applying GSOM to RatSLAM*

Previously, the view cells module of RatSLAM did not impose competition between cells [30], rather a profile based comparison or template matching is performed [10]. In contrast, the proposed framework of the view cells, depicted in Fig. 2, is implemented using GSOM to achieve an incremental scene learning and recognition. The algorithm computes Gist features for each image, whereas the distance of the current feature vector from existing neurons is determined using (5). If a neuron is already familiar with the current place, it would cause that neuron to fire $a_i = 1$ while other neurons would not show a response. A best matching neuron is identified using the proposed objective function (see Section 3.3.3). The activation level is determined from the frequency V_i maintained for each neuron that is updated every time a neuron gets activated. In case of a novel scene, a new neuron is created to learn the pattern on the basis of criteria described in Section 3.3.5. This information is associated with back-end of the RatSLAM system to perform localization and mapping, governed by the pose cells network **P** and experience map **E**. The pose cells network is a 3D grid of a continuous attractor network (CAN) representing the position (x, y, θ) of a robot. The experience map maintains a sequence of experiences with each experience representing the state of a robot i.e., the pose and the associated visual scene (for details see [11, 10]).

4. RESULTS AND EVALUATION

To test the performance of the proposed approach, experiments are performed on the St. Lucia dataset downloaded from the RatSLAM's web page [19]. The resolution of the available video is 640×480 and it contains 2517 frames including the path which has been traveled twice during the phase of data acquisition. Unfortunately, ground truth is not available for the specified dataset, therefore, we had to manually label each frame in the video sequence as visited or unvisited. In this regard, we observed that the vehicle enters the already traveled path after 1735 frames and travels along that path for 30 seconds. Hence, a frame is tagged as a visited place if it is encountered within this time period. The usefulness of the Gist features is tested using a distance matrix, computed for a sequences of 2000 frames, as shown in Fig. 3. The matrix shows the distances among individual scenes, which are computed using distance metric defined earlier in Section 3.3.2. These scenes include the path that has been re-visited by the vehicle. The diagonals appearing on either sides of the main diagonal show the regions of re-visit. This is the road in the dataset which has been traversed twice. It can be observed that Gist features perform well for place recognition and detecting loop closures, despite the false negative recalls which are caused by partially occluded views or considerable changes in the scene.

4.1 Activity in View Cells

We compare our data association module with the existing profile based place recognition module of the RatSLAM. In this regard, the activity in view cells **V** is analyzed and false positive and negative cases are observed for the entire course of mapping. The activity of view cells in our Gist+RatSLAM approach is stable compared to RatSLAM, as shown in Fig. 4. It has also shown the ability to compactly represent the environment using only 297 neurons. In contrary, the RatSLAM approach created 541 visual templates to represent the same environment. This refers to the fact that Gist+RatSLAM imposes less memory demands as opposed to the existing RatSLAM's front-end and thus it can be used for mapping large scale environments. Moreover, it can be seen that RatSLAM produced more false negative recalls when a learned place is re-visited, while Gist+RatSLAM flagged more correct recalls (true positive) on traversal of the learned places. A small area infront of the purple shaded region depicts start of the path which has been driven previously but not detected by either of the algorithms. The purple region in the figure shows the distance in number of frames when a first true positive recall was

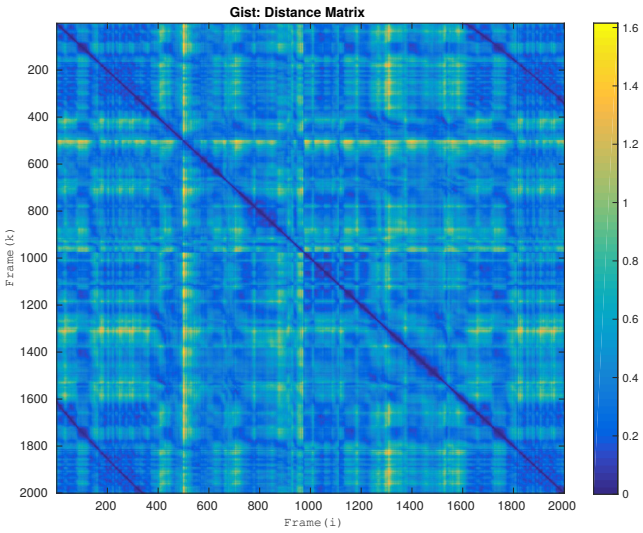

Figure 3: Distance matrix showing similarity among individual scenes. The values close to 0 show that places are closer on the perceptual axis in the space characterized by Gist features, whereas brighter regions depict distant scenes.

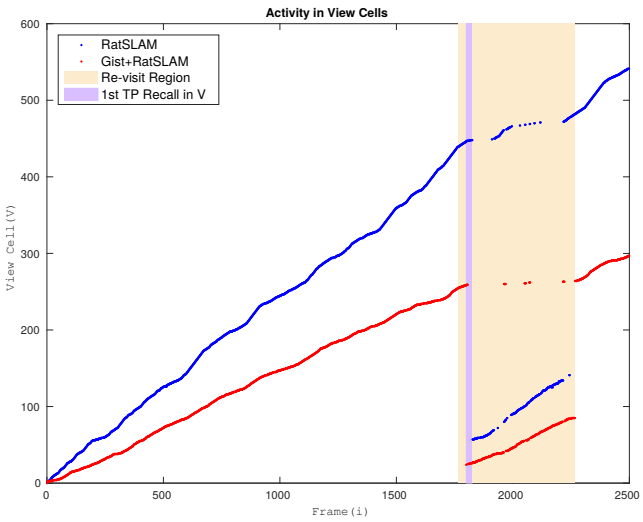

Figure 4: Activity in view cells for traversing unvisited and visited places. Gist+RatSLAM (our approach) needs fewer view cells and has more true positive recalls compared to RatSLAM (blue curve).

responded. This reveals two major findings: first, it reflects that RatSLAM made more incorrect recalls in relation to Gist+RatSLAM; second, the detection of the first true positive recall is very late for RatSLAM and hence it has a slower convergence and loop closure detection compared to our method.

Optical cameras are often susceptible to noise and environmental changes such as particles in the environment, illumination conditions and motion blur due to an inappropriate frame rate. This leads to false positive or false negative results. Hence, we simulated these two scenarios in the St. Lucia dataset to evaluate the performance of our Gist+RatSLAM algorithm under such uncertain conditions.

4.1.1 *Robustness for Gaussian Noise*

To test the strength of our approach, Gaussian noise is added to every input image and then the Gist descriptor is computed. The value for σ of the Gaussian is chosen randomly every time such that $\sigma \in [0.01, 0.1)$ for $\mu = 0$. This kind of corruption is very common during navigation tasks due to the particles present in the atmosphere. We observed that the Gist+RatSLAM remained reasonably tolerant to the noise, because only few places are misclassified while traversing the previously visited path, see Fig. 5. However, this time 345 view cells are created to represent the same environment, which is an obvious response to a noisy input. RatSLAM's profile-based matching created a slightly larger number of view cells i.e., 545 cells as compared to the case when noise was not added. On the other hand, it can clearly be seen that the area before the purple shaded region has been expanded. This indicates that the rate of incorrect recalls has increased for both the algorithms. We have also noticed some false positive recalls in the unvisited region, where both the algorithms falsely associated the current place to the recently viewed scene.

4.1.2 *Performance with Motion Blur*

To simulate motion blur in the dataset, an in-plane motion blur is applied opposite to the direction of the camera rotation. The rotation of the camera is obtained from rough visual odometry computed by the RatSLAM system. A motion-blurred image is one of the common reasons which affects the robustness of many descriptors. For RatSLAM, it caused an increased number of false positives, whereas Gist+RatSLAM is not influenced substantially, as can be observed in Fig. 6. Unlike previous scenarios, in this particular case, the activity in view cells reflects a drop-off in the number of cells utilized to map the entire trajectory traveled by the vehicle. Here, 253 neurons were created for our approach, whereas RatSLAM utilized 372 visual templates to represent the environment. An apparent reason to this phenomena is the fact that motion blur suppresses high frequency components in an image so less information would be available to discriminate between consecutive places. One should note that suppression of high frequencies has no relation to the Gist of a scene, rather it is formed from both high and low frequency components at different spatial scales [18].

4.2 Precision-Recall Rate

The response of Gist+RatSLAM or RatSLAM is regarded as true positive (TP) if it correctly recalls a visited place and it is considered true negative (TN) if the algorithm predicts an unvisited place as a novel place. The false positive (FP) and false negative (FN) responses are incorrect decisions made by the algorithm, i.e., unvisited places are predicted as familiar and vice versa. So, precision-recall is computed as follows:

$$Precision = \frac{\#TP}{\#TP + \#FP}$$

$$Recall = \frac{\#TP}{\#TP + \#FN}$$

Usually, the precision-recall itself is not enough to determine the accuracy of tests, and so the results can be misinterpreted. That is why, the harmonic mean of precision and recall, known as *F1 Score*, is also computed to deduce

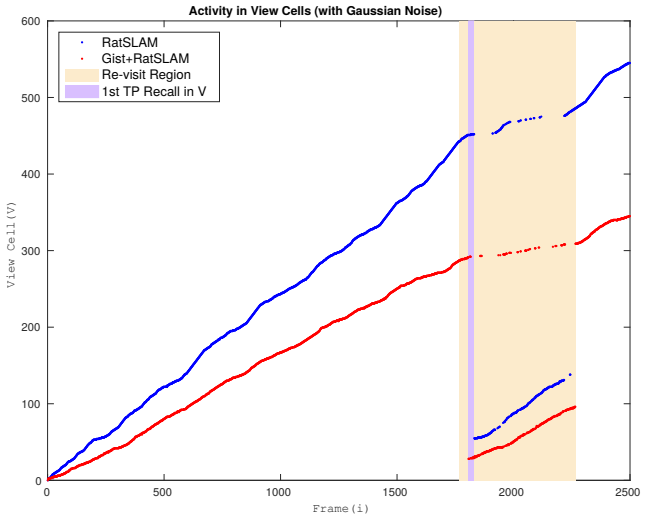

Figure 5: Activity in view cells when Gaussian noise is added to every frame. Gist+RatSLAM is able to detect a loop closure before RatSLAM and shows relatively good convergence.

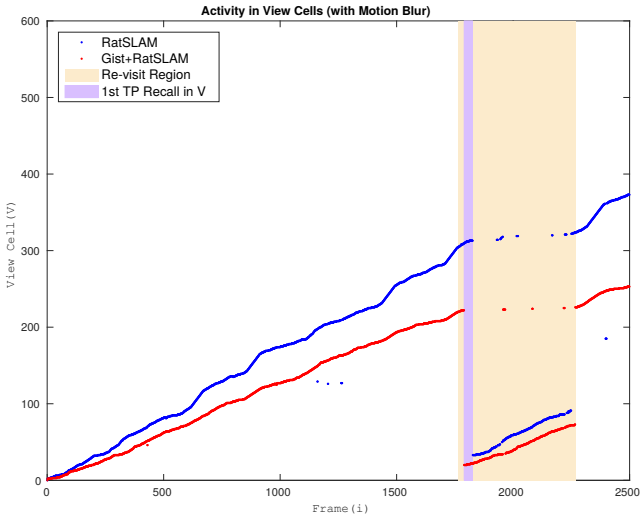

Figure 6: Applying motion blur to images opposite to the direction of rotation induces false positive activity in the view cells for RatSLAM, whereas Gist+RatSLAM shows robustness against outliers.

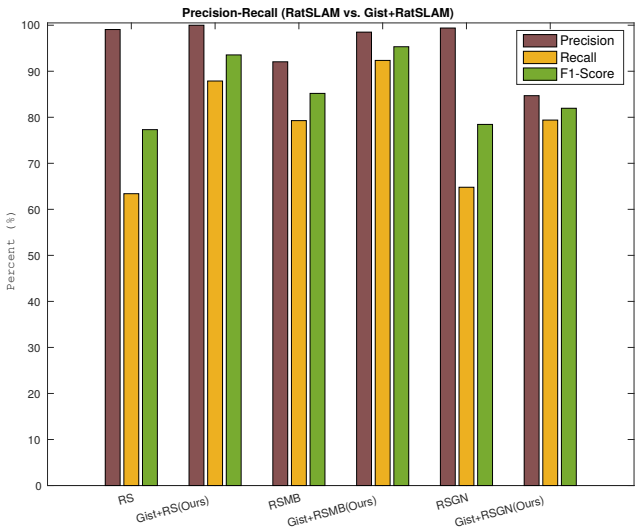

Figure 7: Precision-recall rate for different scenarios: RatSLAM shows 99.39% precision for the additive Gaussian noise case (RSGN). But it has only 64.8% recall-rate compared to our Gist+RSGN indicating 79.4% recall; interpreting results with F1 metric shows robustness of our method for all cases.

inferences about the overall performance of the algorithms:

$$\text{F1 Score} = 2\,\frac{Precision \times Recall}{Precision + Recall}$$

In order to derive conclusions, precision-recall of RatSLAM and Gist+RatSLAM is computed for each of the cases discussed in the previous sub-sections. In the absence of noise and motion blur, precision for RatSLAM (RS) is 99.06% and Gist+RatSLAM (Gist-RS) has 100% precision. A significant difference is noticeable in recall, where our approach reaches 87.88% correct recall of places; whereas RatSLAM has produced more false negatives for the already visited path (recall 63.4%). When Gaussian noise is added to the dataset, RatSLAM (RSGN) gives 99.39% precision and 64.8% recall. In contrast, our method (Gist+RSGN) achieved 84.7% precision and up to 79.4% recall rate. The rational for this decline in Gist+RSGN precision is the fact that Gist is composed of both low and high frequency components calculated at different scales and orientations, while noise affects higher frequencies in an image. One can misapprehend these results to comment the overall performance of the algorithms, thus F1 Score should be used that shows a higher confidence of 81.96% in Gist+RSGN while it is 78.45% for RSGN. With regard to the motion blur case, our approach (Gist+RSMB) yet outperformed RatSLAM (RSMB) having 98.5% precision and 92.35% recall, respectively.

Finally, the maps obtained from Gist+RatSLAM and RatSLAM, respectively, for the driven path are shown in Fig. 8. At a glance it is clear that both algorithms detected the loop closure and preserved the topology of the path traversed by vehicle. However, with reference to Section 4.1, it is worth mentioning that our method shows comparatively a faster convergence. Also, it builds the compact representation of an environment and created less experiences than RatSLAM.

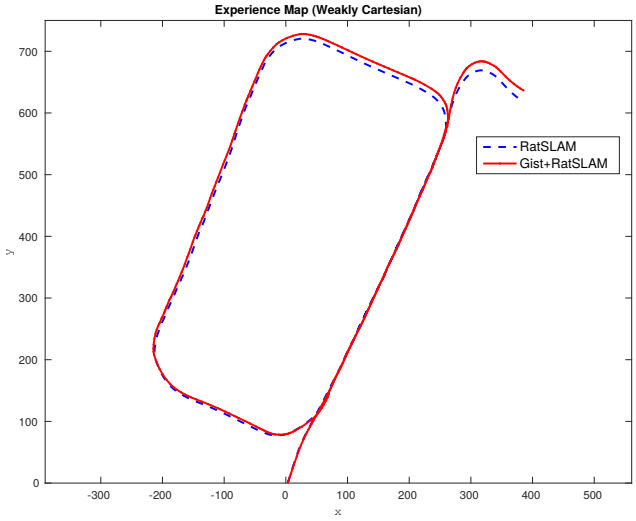

Figure 8: Experience map of the path driven by vehicle. RatSLAM and Gist+RatSLAM are able to detect loop closure and preserve topology of environment, but Gist+RatSLAM has specifically better convergence as already demonstrated in Section 4.1

5. CONCLUSIONS

In this paper, we proposed a bio-inspired front-end that governs the data association for RatSLAM. To accomplish this task, Gist features have been used and a modified growing self-organizing neural network is implemented, which models the competitive behavior of the cells found in visual and perirhinal cortices. This has allowed us to realize online place learning in unknown environments. The results obtained from the experiments on the St. Lucia dataset demonstrate the robustness of our method for noisy and blurred images. We are able to achieve a better recall rate (and thus faster convergence) compared to existing data association module of the RatSLAM algorithm. The ability of the proposed method to compactly represent environments makes this work useful for learning large scale environments. We would be therefore interested to extend and evaluate this work for even larger routes and comparing our approach with state-of-the-art appearance based mapping approaches e.g., FAB-Map that is based on learning local keypoint descriptors. Local descriptors are sensitive to noise and illumination conditions compared to global descriptors (such as Gist features) and demand high computation time. The present research in cognitive psychology suggests that in early vision the human interpretation of a scene is followed by a coarser representation of an environment. It is therefore desirable to extend this work to hierarchical scene learning by combining Gist features with local keypoint descriptors. Moreover, we believe that the current objective function to find the winner neurons can be improved by taking account of factors, such as self-motion cues.

6. REFERENCES

[1] A. Arleo and W. Gerstner. Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. *Biological Cybernetics*, 83(3):287–299, 2000.

[2] Z. Chen, A. Jacobson, U. Erdem, M. E. Hasselmo, and

M. Milford. Towards bio-inspired place recognition over multiple spatial scales. In *Australasian Conference on Robotics and Automation*, 2013.

[3] M. Cummins and P. Newman. Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, 2011.

[4] L. Fausett. *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall, Inc., 1994.

[5] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth. Fab-map+ ratslam: appearance-based slam for multiple times of day. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 3507–3512. IEEE, 2010.

[6] E. Goldstein. *Cognitive psychology: Connecting mind, research and everyday experience*. 2014.

[7] G. Grisetti, R. Kümmerle, C. Stachniss, and W. Burgard. A tutorial on graph-based slam. *Intelligent Transportation Systems Magazine, IEEE*, 2(4):31–43, 2010.

[8] A. Kawewong, N. Tongprasit, S. Tangruamsub, and O. Hasegawa. Online and incremental appearance-based slam in highly dynamic environments. *The International Journal of Robotics Research*, 2010.

[9] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[10] M. Milford and G. Wyeth. Persistent navigation and mapping using a biologically inspired slam system. *The International Journal of Robotics Research*, 29(9):1131–1153, 2010.

[11] M. J. Milford and G. F. Wyeth. Mapping a suburb with a single camera using a biologically inspired slam system. *Robotics, IEEE Transactions on*, 24(5):1038–1053, 2008.

[12] M. Montemerlo and S. Thrun. *FastSLAM: A scalable method for the simultaneous localization and mapping problem in robotics*, volume 27. Springer, 2007.

[13] A. C. Murillo, G. Singh, J. Kosecka, and J. J. Guerrero. Localization in urban environments using a panoramic gist descriptor. *Robotics, IEEE Transactions on*, 29(1):146–160, 2013.

[14] R. Oates, M. Milford, G. Wyeth, G. Kendall, and J. M. Garibaldi. The implementation of a novel, bio-inspired, robotic security system. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 1875–1880. IEEE, 2009.

[15] J. O'Keefe. A review of the hippocampal place cells. *Progress in Neurobiology*, 13(4):419–439, 1979.

[16] A. Oliva and P. G. Schyns. Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, 34:72–107, 1997.

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

[18] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155:23–36, 2006.

[19] RatSLAM. Robotics@QUT-St Lucia Dataset, 2009. https://wiki.qut.edu.au/display/cyphy/RatSLAM [Accessed: 2015-10-25].

[20] K. Rebai, O. Azouaoui, and N. Achour. Bio-inspired visual memory for robot cognitive map building and scene recognition. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 2985–2990. IEEE, 2012.

[21] H. Sasamura and T. Saito. A simple learning algorithm for growing self-organizing maps and its application to the skeletonization. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 1, pages 787–790. IEEE, 2003.

[22] C. Siagian and L. Itti. Biologically inspired mobile robot vision localization. *Robotics, IEEE Transactions on*, 25(4):861–873, 2009.

[23] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[24] N. Sünderhauf and P. Protzel. Beyond ratslam: Improvements to a biologically inspired slam system. In *Emerging Technologies and Factory Automation (ETFA), IEEE Conference on*, pages 1–8. IEEE, 2010.

[25] N. Sünderhauf and P. Protzel. Brief-gist-closing the loop by simple means. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 1234–1241. IEEE, 2011.

[26] A. Tapus and R. Siegwart. A cognitive modeling of space using fingerprints of places for mobile robot navigation. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 1188–1193. IEEE, 2006.

[27] S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics*. MIT press, 2005.

[28] S. Thrun and M. Montemerlo. The graph slam algorithm with applications to large-scale mapping of urban structures. *The International Journal of Robotics Research*, 25(5-6):403–429, 2006.

[29] M. Tomono. 3d localization based on visual odometry and landmark recognition using image edge points. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 5953–5959. IEEE, 2010.

[30] G. Wyeth and M. Milford. Spatial cognition for robots. *Robotics & Automation Magazine, IEEE*, 16(3):24–32, 2009.

[31] Z. Zivkovic, O. Booij, and B. Kröse. From images to rooms. *Robotics and Autonomous Systems*, 55(5):411–418, 2007.