Scaling up a Boltzmann machine model of Hippocampus with visual features for mobile robots

Alan Saul, Tony Prescott, Charles Fox
Sheffield Centre for Robotics
University of Sheffield S10 2TN, UK
aca08ads@shef.ac.uk

***Abstract*— Previous papers [4], [5] have described a detailed mapping between biological hippocampal navigation and a temporal restricted Boltzmann machine [20] with unitary coherent particle filtering. These models have focused on the biological structures and used simplified microworlds in implemented examples. As a first step in scaling the model up towards practical bio-inspired robotic navigation, we present new results with the model applied to real world visual data, though still limited by a discretized configuration space. To extract useful features from visual input we apply the SURF transform followed by a new *lamellae*-based winner-take-all Dentate Gyrus. This new visual processing stream allows the navigation system to function without the need for a simplifying data assumption of the previous models, and brings the hippocampal model closer to being a practical robotic navigation system.**

I. Introduction

The hippocampus is thought to play a key role in associating an observed situation with similar memories [11], [8], [16], [9]. An alternative school of thought is that the hippocampus is a centre for navigation and map building [17], [1], [3]. Both approaches associate its Dentate Gyrus (DG) with the task of pattern separation [2], that is, creating non-overlapping representations of state which can be used to further discriminate between memories. All of these ideas could provide useful inspiration for localisation and mapping in mobile robotics, as explored in the RatSLAM system [12].

There are many models of the hippocampus which are gradually converging toward each other, but this paper will build on the 'unitary coherent particle filter hippocampus' (UCPF-HC) mapping of [4], [5] which begins with an explicitly Bayesian algorithm and works top-down towards the biology. The model has previously been presented primarily as a biological theory, but here we make initial steps towards using it as a practical robotic localisation and mapping method, by extending it to use more realistic sensory inputs. This requires machine vision features and some changes to the biological mapping.

This paper first gives a brief review of the hippocampus, the UCPF-HC model, and of the visual SURF features. It then presents a visual sensory extension to the model including a modified Dentate Gyrus and CA1 to handle the new sensors. We then show that the visual sensors allow for more realistic navigation – though still in an artificially discretized world – as a step towards bio-inspired mobile robotics applications.

A. Anatomy

The principal input structures of the hippocampus are the superficial layers of Entorhinal Cortex (ECs). ECs projects to Dentate Gyrus (DG) which is believed to sparsify the encoding of ECs. Both ECs and DG project to area CA3, which also receives strong recurrent connections that are disabled [7] by septal acetylcholine (ACh). CA3 and ECs project to area CA1, which in turn projects to the deep layers of Entorhinal cortex (ECd), closing a loop if ECd sends information back to ECs. ECs, CA1 and ECd outputs appear to share a coding scheme, as evidenced by one-to-one topographic projections. In contrast, DG and CA3 outputs are thought to work in other bases or latent spaces. In a second loop, ECs and CA1 both project to Subiculum (Sub), which projects to the midbrain Septum (Sep) via fornix. Septal ACh and GABA fibres project back to all parts of hippocampus.

B. UCPF-HC model

The UCPF-HC model [4], [5] mapped this hippocampal circuit onto a modified Temporal Restricted Boltzmann machine (TRBM, [20]), a machine learning algorithm. The TRBM is a Bayesian filter with Boolean observation vectors (including an always-on bias node), z'; Boolean hidden state vectors (also including an always-on bias node), x'; weight matrices $W_{x'z'}$ and $W_{x'x'}$. It specifies joint distributions,

$$P(x_t, x_{t-1}, z_t) = \frac{1}{Z} \exp \sum_t (-x'_t W_{x'x'} x'_{t-1} - x'_t W_{x'z'} z'_t). \tag{1}$$

Unlike the standard TRBM, the unitary coherent particle filter hippocampus mapping uses the following deterministic update to obtain maximum *a posteriori* estimates:

$$\hat{x}_t \leftarrow \arg\max P(x_t|\hat{x}_{t-1}, z_t) \tag{2}$$
$$= \{\hat{x}_t(i) = (P(x_t(i)|\hat{x}_{t-1}, z_t) > \frac{1}{2})\}_i$$

which is the zero-temperature limit of an annealed sequential Gibbs sampler. A version of the wake-sleep algorithm is mapped onto particular phases of the hippocampal theta cycle in [4], conjecturing use of after-depolarization effects to reset the wake-sleep stages.

Noisy inputs $z_t = y_t + \epsilon_t$ are mapped to the combined ECs and DG, where the DG activations are functions of the ECs activations, $z_t = (ECs_t, DG_t(ECs_t))$. CA3 is mapped to

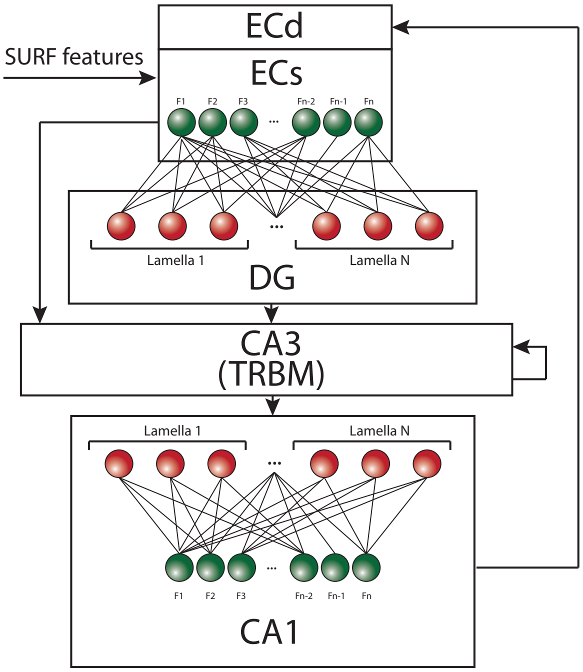

Fig. 1: Illustration of the hippocampus model showing data flows and hippocampus regions. SURF features are the new visual inputs. (Subiculum circuit not shown.)

the hidden state, x_t. CA1 performs a partial decoding into the DG basis. Finally the estimated de-noised output is mapped to ECd, $\hat{y}_t = ECd_t$. Each neural population is a Boolean vector at each discrete time step t.

A major problem with UCPF-HC is tracking loss, as it approximates whole posteriors with single samples. To recover from tracking loss, filter performance is monitored to heuristically detect its occurrence – by thresholding a moving average of discrepancy between observed and denoised sensors – then the priors are disabled when lostness is detected. In UCPF-HC, the Subiculum-Septum circuit performs this monitoring. Subiculum then compares the partially decoded CA1 information against the original ECs input, receiving one-to-one connections from both regions. If they differ for an extended period of time, this indicates loss of tracking, and the recurrent CA3 connections are disabled by Septal ACh.

The UCPF-HC model was shown in [5] to learn receptive fields of CA3 cells corresponding to a mixture of places and world features, as found in biology. This is in contrast to pure place-cell models which can learn to perform mapping and localisation (SLAM, e.g. [12]) but do not exhibit the world feature detectors found in the biological primate CA3 [11], [8], [16], [9].

C. *Visual SURF features*

The UCPF-HC model was originally intended as an explanatory theory of the biological hippocampus, though a simple proof of concept computational implementation was provided, running in a highly simplified microworld. This included highly abstract touch and vision senses, but the present paper will use a more advanced and realistic visual feature input, SURF.

Speeded-Up Robust Features (SURF) [6] are a state-of-the-art transform from images to a vector feature space, designed to be informative for recognising objects in machine vision. They have also been found useful in other navigation algorithms, and are an evolution of the older Scale Invariant Feature Transform [10]. The SURF transform begins by detecting interest points in an image, found as the local maxima of the scale pyramid of the image convolved with Haar wavelets. The dominant wavelet orientation is found at each interest point. Using the found interest point and orientation, a small localised grid is constructed, and sums of absolute and signed responses of vertical and horizontal Haar wavelets in its cells taken as a 64-element feature vector.

II. Methods

A. *Environment and sensors*

The microdomain used in the previous UCPF-HC models was a simulated plus-maze environment, containing 13 discrete locations consisting of four arms each containing 3 discrete locations, and a centre point. The robots state within this environment is encoded by one of 13 places ($place \in [0:13]$) and one of 4 head-directions corresponding to discrete compass headings ($hd \in \{N, E, S, W\}$). Touch sensors detect the presence of surrounding walls, and cells responding to coloured posters at the ends of arms simulated a crude form of colour detection.

Previously this simulated environment was completely theoretical and senses were crafted to allow for maximal separation between locations descriptions. To bring the model closer to a functional robotic application, a representation of vision has been introduced to the original model's array of senses. Each location direction pair, (p, hd), has a selection of associated views. Images correspond to real photographs taken from within the courtyard of Regent Court, University of Sheffield (Figure 2), whose paved paths form a real-world plus maze shape. Photographs from the real location are directly mapped to the simulated environment. Alternative views of the same location are used to simulate changes in the environment and incorrect odometry input upon revisiting a location. Alternative images were taken at different times of day, at slightly differing angles ($\pm 10°$) under different lighting and weather conditions. To overcome the effective noise introduced by using images under different lighting conditions, robust SURF features are used to encode the significant features of each image and detect similarities between new and previously seen views.

SURF features are extracted for every image, and are compared with a base set of SURF features. If the SURF feature is found to be a close enough match to any feature in the base set, the base set feature is determined to be present for that image. This allows a Boolean visual observation vector to be formed uniquely describing an image as a combination of present features. This visual observation vector is fed into the hippocampal model via the EC input

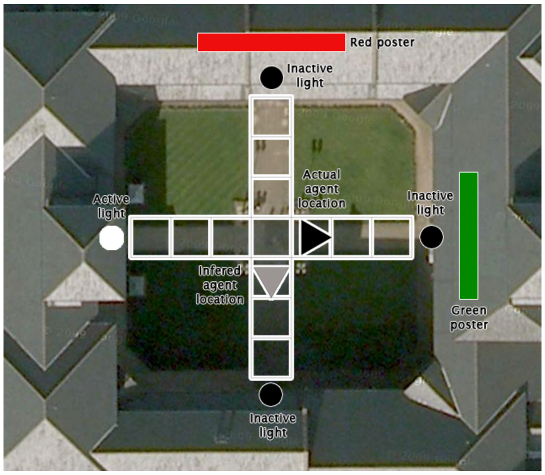

Fig. 2: An overhead view of Regent Court (Google Maps), showing the mapping of the artificial plus maze used in simulation to real life location.

and is processed alongside the existing odometry and sensory information throughout learning and inference.

B. Pre-processing visual features

The base set of features required for matching must be a range of features which are present in a range of images. In order to extract SURF features that are present in a range of images, the descriptions of individual SURF features need to be merged. SURF features produce a description of a feature in the form of 64 floating point numbers. The approach used to extract common SURF features that are as closely matched as possible by a range of images, was to use a combination of the *k*-Nearest-Neighbour algorithm and merging. The merged SURF features produced are similar to that of merging groups using *k*-Means-Clustering and taking the average SURF feature description for each group. The distance between a pair of SURF features $\mathbf{p}^i$ can be determined as the Euclidean distance between the two features,

$$d(\mathbf{p}^i) = \sum_{j}^{64} (\mathbf{p}_j^{i[1]} - \mathbf{p}_j^{i[2]})^2. \quad (3)$$

Algorithm 1 describes in depth the how this generalisation was made, accompanied by Figure 3. The Fast Library for Approximated Nearest Neighbour search [13] was used for find the nearest neighbour with $k = 2$.

Determining whether a feature is present within an image requires taking the Euclidean distance of Haar wavelet responses between a member of the merged SURF feature set, and a SURF feature extracted from the image being analysed (Equation 3). A feature is regarded present if this distance is small enough, presence is thus a function of distance between Haar wavelet responses,

$$c(\mathbf{fe_j}, \mathbf{fe_k}) = \begin{cases} 1 & \text{if } d(\mathbf{p}^i) < t' \text{ s.t } \binom{p^{i[1]}=\mathbf{fe_j}}{p^{i[2]}=\mathbf{fe_k}} \\ 0 & \text{otherwise} \end{cases} . \quad (4)$$

Algorithm 1 SURF feature generalisation

```
Extract most distinctive features from every image
Plot these features in 64 dimensional space using their descriptors
Use $k$NN to find nearest neighbour to every individual feature
while feature pair $\mathbf{p}$ exists s.t $d(\mathbf{p}) < t$ do
  for each pair: $\mathbf{p}^i$ in P s.t $d(\mathbf{p}^i) < t$ do
    Merge the two features Haar wavelet responses:
      $\mathbf{p}^{n+1} = \frac{\mathbf{p}^{i[1]}+\mathbf{p}^{i[2]}}{2}$
    Add $\mathbf{p^{n+1}}$ to feature set
    Remove $\mathbf{p}^{i[1]}$ and $\mathbf{p}^{i[2]}$ from feature set
  end for
end while
```

The maximum distance to be regarded as present, t', depends on the particular set of merged features and the features being analysed. It is the case, however, that the distance from the original features to the newly merged features is likely to be larger than the t used for merging (Algorithm 1), if the feature has been used in multiple merges. In order for even the same features that were used in the initial merging to be present, the similarity threshold between features required to indicate a 'match' needs to be relaxed to t' such that $t' > t$. Without merging features and relaxing the matching threshold t', images will share very few common features. Consequently a large Boolean observational vector would be required to ensure each image could be uniquely described by its combination of SURF feature matches.

C. DG encoding

The previous UCPF-HC models used a small number of simple touch and colour sensors, along with grid cells, as the ECs input vector. Because these sensors were well understood by the model authors, it was possible to handset weights $W_{EC \to DG}$ to make DG cells respond to particular combinations of inputs, such as touches and locations.

The new model retains these handset weights for the simple sensors, but as it introduces a large number (e.g. 80) of merged SURF features into the ECs input, it is no longer practical to choose and hand-set useful combinations of these additional features, and so an automated approach must be used instead.

The *lamellae hypothesis* [18], [14], [19] in neuroscience states that single cells in EC project exclusively to localised (2mm) regions of DG, known as *lamellae*. Following [14] we model the visual part (only) of our DG as a set of N lamellae, receiving input from subsets X of the merged SURF features only.[1]

The Hebbian, winner-take-all learning of Algorithm 2 is applied within each lamella, which forces it to sparsely

[1] In the full lammelae hypothesis, each lamella receives input from many types of EC cells such as touch, grid and vision. However we chose not to alter the structure of the old model's sensor processing, in order to keep the visual extensions as a removable/pluggable extension to the model.

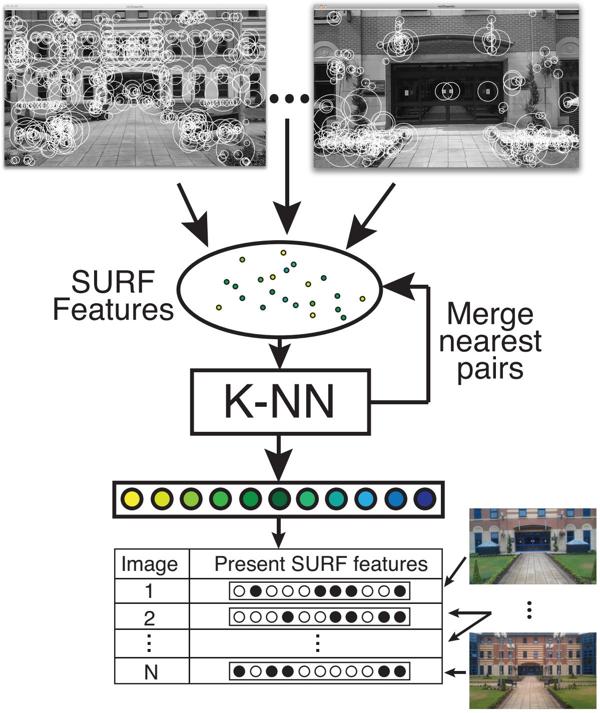

Fig. 3: Illustration of the process required to derive a set of SURF features common to multiple images, and computing SURF description of images.

encode its input. Consequently only one neuron in each lamella will be active at any time, and each lamella learns to represent a set of mutually exclusive conjunctions of its inputs. This ensures that the DG activity maintains sparseness, as seen in biology, with only N neurons being active in the entire DG subfield of $N \times X$ neurons.

D. CA3, CA1, ECd decoding

CA3 projects to CA1, which in turn projects back onto the deep layers of the EC (ECd). Several computational models [15], [4] view the CA1 as a translator, helping to map from the CA3 coding scheme back to the ECd representation. As in the previous UCPF-HC model [4], we assume that CA1 decodes the output of CA3 into the same coding scheme used by DG, then ECd decodes from this to the same scheme used by ECs.

As in the previous model, we do not attempt to learn the weights $W_{CA3 \to CA1}$ in a biologically plausible way[2]. Rather, we simply re-use the weights $W_{DG \to CA3}$ learned by the TRBM. A naïve decoding would be given by

$$CA1_{naive} = (sig(W^T_{DG \to CA3} CA3) > \frac{1}{2}), \qquad (5)$$

and ignoring the EC components of the resulting (EC,DG) vector. This decodes each CA1 cell individually using a threshold. However, a better decoding is possible because we have prior knowledge about the structure of DG's lamellae, which is reproduced in CA1. We know that due to winner-take-all encoding, exactly one cell of each lamella will be active. Hence we may apply the winner-take-all rule again in each CA1 lamella[3], which has the effect of denoising the CA1 in comparison to the naïve method.

Algorithm 2 Competitive learning algorithm for N neural networks

```
for all lamella $n$ in $N$ do
    Choose $X$ random feature indices
    Make fully connected neural network connecting EC[X] features to DG lamella $n$
    Randomise weights of all $W^n = W_{EC[X] \to DG[n]}$
end for
for $P$ presentations of training data do
    for all image $I$ in training set do
        for all lamella $n$ in $N$ do
            Calculate output firing:
                $O^n = I \cdot W^n$
            Find winning neuron in lamella $n$:
                $a = \max(O^n)$
            Update weights:
                $\Delta W^n_{ij} = \alpha O^n_j I_i$ if $j = a$ else 0
            Normalise weights:
                $\mathbf{W^n} = \frac{\mathbf{W^n}}{\sum_{ij} W^n_{ij}}$
        end for
    end for
end for
```

The weights $W_{CA1 \to ECd}$ need to invert the previous sparse coding $W_{ECs \to DG}$ to return information to the original entorhinal input basis. In the new model we have been able to do this in quasi-biological way, using the classical perceptron learning rule of algorithm 3.

Algorithm 3 Perceptron learning rule

```
Initialise weights to zero.
for $P$ presentations of training data do
    for all lamellae $n$ in $N$ do
        for all encoded patterns $ideal$ of input patterns $I$ do
            Calculate actual output activity of CA1:
                $y_{Ideal} = [f(\mathbf{w}^n \cdot Ideal)]$
                where $f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$
            Update weights:
                $\Delta \mathbf{w^n} = \alpha(I - y_{Ideal}) ideal$
        end for
    end for
end for
```

III. Results

The present model extends the previous UCPF-HC model [4] by adding realistic visual input and new DG and CA1 processing of that visual input, after SURF preprocessing. It retains the discrete locations of the plus maze from the old model, but places them into real spaces around the Regent

[2] though see [4] for a discussion of how it could be achieved.

[3] in our code this is described as 'smart decoding'.

Court building as described in section II-A. One would expect the introduction of additional sensors and processing to improve the performance of the model.

We test the new model using the same protocol as in the previous paper [4]. 30,000 random walk steps are taken around the 13 discrete locations of the simulated Regent Court environment, and the walk was replayed though the learning algorithms until the weights converged. We used a number of input neurons feeding into each lamellae, $N = 7$ and a number of lamellae, $X = 45$, throughout all experiments. To simulate noise within the odometry and sensory input, the EC had 10% noise, ie. on average one in 10 of its Boolean senses flipped. Python code is again available as supplemental material from the authors.

The results show that the additional visual information produces dramatic improvement in the model's ability to maintain an estimate of its true location.

Figure 4 shows the amount of time that the agent is lost during the walk, (following the display format of [5], [4]) in runs with the new SURF extensions disabled and enabled. In a previous test of the learning model [5], a simplifying assumption was made[4] that the grid cells behave as 'noisy GPS' [4] units rather than accumulating odometry data – this was to make the learning problem easier while demonstrating the biological learning mapping. Figure 4 now shows the results of the original UCPF-HC model without this assumption, *No SURF No nGPS*, which are in fact very poor and comparable to *Random* weights[5]. The figure reproduces previous results [4] for performance with the learned model with the nGPS assumption made, *Learned with nGPS*.

The two bars on the right of figure 4 show the performance of the new model, with the new visual system enabled. Using SURF features (and odometry) only, and without the nGPS assumption, the model (*Learned with SURF only*) outperforms the old model, achieving 95% accuracy. When the old model's sensors are included in addition to SURF (*Learned with SURF*), the accuracy improves further to 97%.

Figure 5 shows the lostness probabilities with a larger, 20%, noise introduced to the input sensory information. This shows that extremely noisy data still cannot be handled by the present model thus there is room for improvement in these cases, for example by using further sensors as well as vision.

Figure 6 shows the real location of the robot (black line) against the estimated location by the hippocampal model within (blue line). These lines overlap throughout the whole simulation, showing that a highly accurate estimation is maintained. (See [5] for description of the complex display format.)

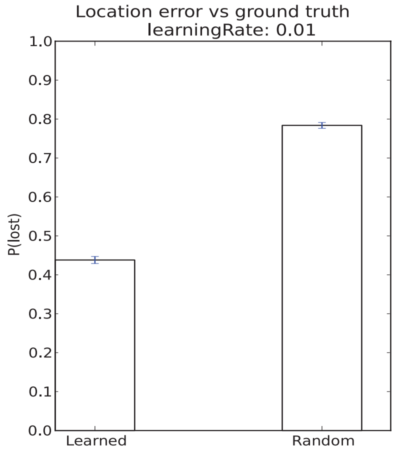

Fig. 5: Lostness probabilities working with 20% noisy data input to EC.

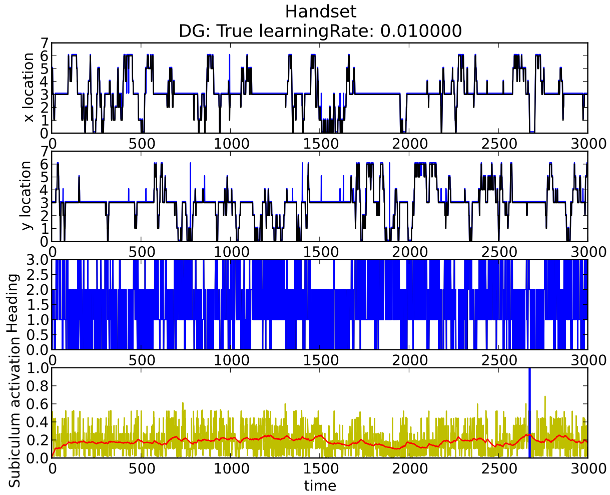

Fig. 6: Results showing almost perfect mapping between ground truths of location and belief of location illustrating extended models ability, with the septum intact.

IV. Discussion and Future Work

Previously the UCPF-HC model [4] used odometry and simplified abstract sensory information. This input included unrealistic 'noisy GPS' grid cell activity, whisker touch senses and extremely simplified colour senses, with handset receptive fields describing descriptive conjunctions of such features in the DG and CA1.

The present study has extended this model to: receive real-world visual SURF features as input; learn its own DG receptive field for these additional features using the biological hypothesis of DG lamellae; learn an improved CA3 representation using this information; and decode it back to ECd using a further lamellae based scheme in CA1.

The visual extensions allow navigation to be performed with high accuracy, even when the noisy GPS assumption is dropped, i.e. when the entorhinal grid cells are performing realistic odometric integration as in a real mobile robot, rather than acting as temporally independent location observations.

[4] documented in its accompanying source code.

[5] In fact they are slightly worse, as the random weights tend to produce very little change in the initial location. A stopped clock is accurate more often than a slow one!

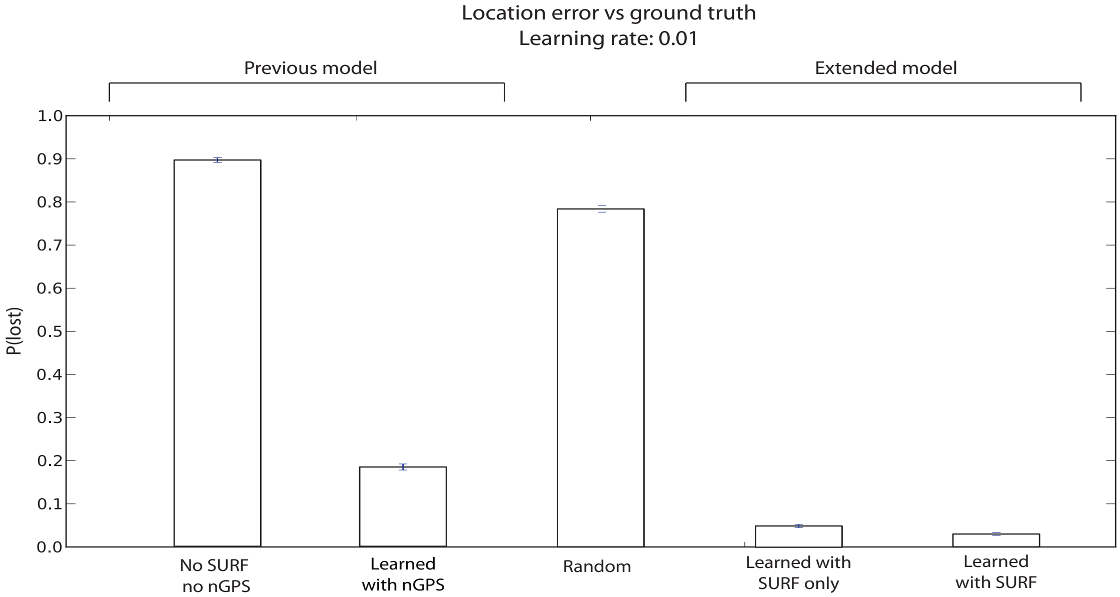

Fig. 4: Comparing lostness probabilities of model containing SURF features and model using only whisker senses, light senses and odometry information.

This level of accuracy suggests that visual SURF features coupled with the UCPF-HC model could form the basis of a future localisation and mapping (SLAM) system for real mobile robots.

However we have still retained the discrete location assumption in this work, which precludes real robot implementation at this stage. The next future work step would thus be to remove this, and allow a real robot using SURF features to learn its own place representations in a continuous world, as performed in RatSLAM [12] and similar architectures. The present work suggests that SURF vision features could be powerful enough to enable this research to take place. In contrast to RatSLAM-like architectures, UCPF-HC is derived from a top-down Bayesian machine learning model, which is able to represent more complex states of the world than place alone, and provides a probabilistic semantic interpretation of the hippocampal function.

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] A Bakker, C B Kirwan, M Miller, and C E L Stark. Pattern Separation in the Human Hippocampal CA3 and Dentate Gyrus. *Science*, 319(5870):1640–1642, March 2008.

[3] R Chavarriaga, T Strösslin, D Sheynikhovich, and W Gerstner. A computational model of parallel navigation systems in rodents. In *Neuroinformatics*, pages 223–241, Ecole Polytech Fed Lausanne, Sch Comp & Commun Sci, CH-1015 Lausanne, Switzerland, 2005. Ecole Polytech Fed Lausanne, Sch Comp & Commun Sci, CH-1015 Lausanne, Switzerland.

[4] C Fox and T Prescott. Hippocampus as unitary coherent particle filter. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–8, 2010.

[5] C Fox and T Prescott. Learning in a unitary coherent hippocampus. *Artificial Neural Networks–ICANN 2010*, 2010.

[6] Bay H, Ess A, Tuytelaars T, and Gool LV. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[7] ME Hasselmo, E Schnell, and E Barkai. Dynamics of Learning and Recall at Excitatory Recurrent Synapses and Cholinergic Modulation in Rat Hippocampal Region Ca3. *J. Neurosci*, 15(7):5249–5262, 1995.

[8] JJ Hopfield. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, 79(8):2554–2558, 1982.

[9] S Kali and P Dayan. The involvement of recurrent connections in area CA3 in establishing the properties of place fields: a model. *J. Neurosci*, 20(19):7463–7477, 2000.

[10] DG Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[11] D Marr. Simple Memory - Theory for Archicortex. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 262(841):23–81, 1971.

[12] MJ Milford, GF Wyeth, and D Prasser. RatSLAM: a hippocampal model for simultaneous localization and mapping. In *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, pages 403–408, 2004.

[13] M Muja. Fast approximate nearest neighbors with automatic algorithm configuration. *International Conference on Computer Vision Theory And Applications*, pages 331–340, 2009.

[14] CE Myers and HE Scharfman. A Role for Hilar Cells in Pattern Separation in the Dentate Gyrus: A Computational Approach. *Hippocampus*, 19(4):321–337, 2009.

[15] RC O'Reilly. Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus, New York, Churchill Livingstone*, 4(6):661–682, 1994.

[16] ET Rolls, SM Stringer, and TP Trappenberg. A unified model of spatial and episodic memory. *Proceedings of the Royal Society of London Series B-Biological Sciences*, 269(1496):1087–1093, 2002.

[17] PE Sharp. Computer-Simulation of Hippocampal Place Cells. *Psychobiology*, 19(2):103–115, 1991.

[18] Lomo T. Excitability Changes Within Transverse Lamellae of Dentate Granule Cells and Their Longitudinal Spread Following Orthodromic or Antidromic Activation. *Hippocampus*, 19(7):633–648, 2009.

[19] N Tamamaki and Y Nojyo. Projection of the entorhinal layer II neurons in the rat as revealed by intracellular pressure-injection of neurobiotin. *Hippocampus*, 3:471–480, 1993.

[20] GW Taylor and GE Hinton. Modeling human motion using binary latent variables. *Advances in Neural Information Processing Systems*, 19:1345–1352, 2007.