Spatio-Temporal Sequence Learning of Visual Place Cells for Robotic Navigation

Vu Anh Nguyen, *Student Member, IEEE* Janusz A. Starzyk, *Senior Member, IEEE*, Alex Leng Phuan Tay, *Member, IEEE* and Wooi-Boon Goh, *Member, IEEE*

Abstract-In this paper, we present a novel biologicallyinspired spatio-temporal sequence learning architecture of visual place cells to leverage autonomous navigation. The construction of the place cells originates from the well-known architecture of Hubel and Wiesel to develop simple to complex features in ventral stream of the human brain. To characterize the contribution of each feature towards scene localization, we propose a novel significance analysis based on the activation profiles of features throughout the spatio-temporal domain. The K-iteration Fast Learning Neural Network (KFLANN) is then used as a Short-Term Memory (STM) mechanism to construct our sequence elements. Subsequently, each sequence is built and stored as a Long-Term Memory (LTM) cell via a oneshot learning mechanism. We also propose a novel algorithm for sequence recognition based on the LTM organization. The efficiency and efficacy of the architecture are evaluated with the vision dataset from ImageCLEF 2010 Competition.

Index Terms— Hierarchical memory architecture, Hubel and Wiesel's model, KFLANN, Spatio-Temporal Sequence Learning

I. INTRODUCTION

Machine intelligence in autonomous navigation concerns mainly two general questions: Localization ("Where am I?") and Mapping ("Where and how do other places relate to me?"). Considering the exploratory task of a target environment, the first problem concentrates on recognizing locations that identify the visited places, while the second problem focuses on representing and self-organizing new locations in memory to build a map of familiar places.

In this paper, we aim at investigating the efficiency and efficacy of a hippocampal-inspired visual place cell model and its spatio-temporal sequence learning to leverage autonomous navigation. The visual place cells characterize the configurations of local appearances across both spatial and temporal domains. The representation of each cell consists of a global or gist feature vector that encodes a visual scene [1]. Each element in the vector corresponds to a local and invariant feature within the visual field. Additionally, the identity of each element is built directly from experiencing the environment. The vocabulary of local elements is constructed based on the feedforward hierarchical architecture of building features from simple to complex with increasing spatial invariance proposed by Hubel and Wiesel [2] and computational models by Fukushima [3], Serre et al [4]. In this work, we also introduce an efficient significance analysis

scheme to characterize and identify the significant local features which contribute mostly to the place identification task. The estimation of feature significance originates from the activation profile of a feature throughout its temporal domain.

An ART-based learning architecture, KFLANN [5], was employed to establish scene STM clusters by global gist description which mimics the fast-learning behavior of scene tokens. The reason for this clustering stage is to maintain a significant tolerance that reflects variations in the explored area. At the same time, it is impractical and not useful to remember all locations within the environment because of limitation of memory capacity and search time requirement. The internal structure of KFLANN is driven by intrinsic statistics based on the data stream. Additionally, the data presentation sequence syndrome in which the clusters set changes with different data ordering is alleviated by an efficient reshuffling mechanism to preserve centroids stability and consistency. These characteristics are critical to reliable sequence identification against various perceptual fluctuations.

Topological structure of the environment is constructed by self-organizing and linking the proposed place cells into temporally ordered sequences of events that compose location-specific episodes. In this work, we propose a novel biologically inspired sequence learning architecture to organize generic scene clusters generated by KFLANN into stable spatio-temporal sequences. We extend the idea of connectionist Long Term Memory (LTM) model in [6], [7] to real-time analog inputs to facilitate navigation. For localization, we will show that by exploiting the sequential properties, the system is able to alleviate ambiguities and enhance reliability in place recognition. This characteristic is useful in recognizing confounding places in which scenes in different places are partially similar. In our model, each sequence of navigating scenes is stored in a LTM cell and is learnt via one-shot mechanism initially. Our sequence recognition algorithm can distinguish among different sequences, as well as is resilient to deviations from ideal sequences. During storage phase, the input sequences are stored in the corresponding LTM cells. During testing phase, the LTM cell will respond according to its degree of matching with the input sequence. The final decision's location is made by the Winner-Take-All (WTA) rule over all LTM cells. Our matching algorithm is also able to work with continuous input sequences in which the beginning or ending point is not specified.

V. A. Nguyen, Alex L.P. Tay and W.B. Goh are with the School of Computer Engineering, Nanyang Technological University, Singapore.

Janusz A. Starzyk is with the School of Electrical Engineering and Computer Science, Ohio University, Athens, USA



Fig. 1. System Architecture

This work serves as an initial investigation on developing hierarchical episodic memory architecture by analyzing the interplay between STM and LTM mechanisms driven by experiences in embodied intelligence. The framework is useful for leveraging navigation performance by exploiting the reliability in sequence of perceptions. The whole system architecture is depicted in Figure 1.

The structure of the paper is as follows: Section II gives the related works to this model. Section III describes in detail the architecture of the hierarchical feature extraction. Section IV follows by introducing neural network for scene clustering by KFLANN network and the proposed sequence learning algorithm. Section V presents some experiment results and analysis. Conclusions as well as future directions are given in Section VI.

II. RELATED WORKS

One of the main advantages of bio-mimetic over probabilistic-based navigational systems is the flexible representation of the target environment. While probabilistic models aim at constructing a high-precision metric map, biological systems evolve to adaptively interact with the environment. The mechanism is conducted via observation/feedback cycle and self-organization into coarse environment-adapted units called place cell [8]. The firing pattern of these cells strongly correlates with particular locations within the environment. Extensive anatomical and psychophysical studies confirm the existence of place cells in hippocampal systems of the brain such as Dentate Gyrus, CA3 and CA1 (see [9] for a comprehensive review). In this work, we model the representation of visual place cells and its learning capability towards scene understanding.

The representation of place cells in our model aims at capturing the holistic structures of the environment [10]. The prominent works typically consider the whole image as a context frames [11] or divide the visual field into smaller grids at fixed positions and sizes [10], [12]. Subsequently, low-level local features, e.g. edges, corners, colors, intensity,

textons at some spatial scales in spatial domain or frequency components in frequency domain are extracted and pooled together to form the global feature vector. The number of dimensions may be further reduced using standard techniques such as Principle Component Analysis (PCA).

In our opinion, the spatial configuration of features is not necessarily analyzed only as groups of features at fixed positions of the visual field but more generally as common patterns of feature-activated locations with possibly some degrees of positional tolerance. Secondly, the lowlevel features should be considered at various degrees of complexity. Ullman [13] studied a wide range of visual features with different complexities and suggested that the intermediately complex features contribute most significantly to classification performance. Thirdly, their contents and scales are not necessarily universal and should be learned from experiences. This requires a systematic way to derive a suitable collection of features that emerge from correlations of visual stimulus.

The development of the proposed place cell model is consistent with evidences from a number of context-sensitive areas of the brain such as regions in Para-hippocampal Cortex (PHC) called Para-hippocampal Place Areas (PPA) and Retrospenial Cortex (RSC) [14]. The fMRI studies show that these areas respond more strongly to pictures which contain scenic structure than to objects alone. To our account, this behavior is strongly related to the prominent properties of place cells in hippocampal systems [9]. Bar [1] shows that PHC and RSC may also associate with characteristics of episodic memory as well as navigation in which cell activations may provide a set of expectations that can guide the perceptions/actions and may influence exploration.

The construction of place cell representation in our place cell model follows the well-known feedforward hierarchical architecture by Hubel and Wiesel [2]. Starting from the visual stimulus at the input, the basic processing stream comprises of consecutive connections of interleaving simple cell (Slayer) and complex cell (C layer) layers with increasing spatial invariance in positions, scales, polarities and orientations following the hierarchy. The hierarchical processing stream involves various cortical regions from LGN, V1 to V2, V4 cortical areas and higher areas of IT cortex [15]. The spatial relationship among complex cells in C layer is preserved at intermediate layers [13]. This preservation is critical for the analysis of scene configuration which comprises of distinct local elements that have high spatial relationship. Thus, we accumulate features at intermediate levels of the hierarchy and use them for our scene analysis. This type of architecture can be dated back to the Neo-cognitron model by Fukushima [3], Convolutional Neural Network model by LeCun [16] and recently HMAX model by Riesenhuber and Poggio [17].

In this paper, we address the important roles of hierarchical memory architecture that adapts and links the spatial episodes of visual place cells into temporal sequences [18]. Functionally, the emerging place cells correspond to STM cells and the spatio-temporal sequence learning of episodes corresponds to activation of LTM cells. The initial experiences are stored in the STM, and then gradually consolidated and organized into LTM. The STM may operate at fastlearning mode to attend to all informative input. However it may suffer from decaying activation. A well-known class for STM mechanism is the Adaptive Resonance Theory (ART) [19] Network. The LTM may operate at slower rate with stable and consistent sequences due to its large highly plastic connection. The key properties of sequence learning models of LTM cells were introduced in series of works by Wang et al [20], [21], [22].

Our previous model in [7] characterized several prominent characteristics of sequence learning such as hierarchical organization, anticipation, and one-shot learning. Subsequent extension in [6] improved the original model by introducing the flexible matching mechanism that gives the real-value degree of similarity between sequences of characters instead of the precise match-nonmatch return. Therefore, it enhances the error tolerance capabilities for distorted, delayed or imperfect starting or ending of sequence. Although the content of visual input used in our model provides rich information and is important to human navigation, it also possesses a large uncertainty due to variations of robot poses and movements that make it difficult for one-shot scene classification by individual data. Exploiting the stability in sequences of observations is useful for tackling this issue. Our model also accepts continuous input stream in order to identify place in real-time manner, giving it the capability to overcome the constraint of imprecise starting and ending points of a given sequence.

III. FEATURE BULDING AND EXTRACTION

Feature analysis can be divided into two main stages: feature vocabulary building and feature extraction. The first stage is to construct a hierarchical architecture of interleaving S layers and C layers while the second stage uses this architecture to extract spatially invariant features and fetch into the LTM sequence learning module.

A. Feature Vocabulary Building

For each input image, low-level features are extracted into several feature maps at S1 layer, each of which results from response of the 2D Gabor filter banks with $n_{\mathcal{O}}$ orientations and n_{S} scales (n_{S} is even), resulting in $n_{O} * n_{O}$ feature maps at S1. Subsequently, all feature maps from layer S1are pooled together to establish complex units at layer C1. Each complex cell in layer C_1 combines a local rectangular group of simple cells with n_{GS} different grid sizes within each S1 feature map and over two S1 feature maps in consecutive scales at the same location. These combinations are conducted separately with each orientation. The activation of each complex cell in layer C1 is the maximum activation of all the simple cells within its receptive field. A group of C1feature maps which result from the sub-sampling by a same grid size is termed a C1 band. Therefore, a total number of $n_{\mathcal{B}}$ ($n_{\mathcal{B}} = n_{\mathcal{S}}/2$) C1 bands are generated for each input image (with $n_{\mathcal{O}}$ orientations).

The construction of the next S2 layer is followed by sampling a large number of rectangular groups of cells across all C1 feature maps in random positions and sizes to develop our S2 cells. For each input image, a number of $n_{\mathcal{P}}$ patches of different size $N_i^{\mathcal{P}}(i=1\ldots n_{\mathcal{P}})$ randomly extracted from all C1 feature maps are used to construct S2 layer. Each C1patch is also associated with the local region of interest E_{P_i} of the size $\tau \cdot N_i^{\mathcal{P}}(\tau = 1.5 \text{ in our experiments})$ centered at its extracted location (x_i, y_i) . For N images, a total number of $N \cdot n_{\mathcal{P}} S2$ feature patches are extracted after this stage. Subsequencely, N_f S2 feature patches are randomly selected from this large collection for the feature extraction stage. After all S2 patches are collected, each of them is connected to a single complex cell in the next layer C2. Therefore, any new image can be represented by a vector of the N_f C2 complex features. The detail of C2 cells' activation computation is described in the next section.

B. Feature Extraction

For an input image, S1 and C1 feature maps are generated as in the previous phase. This operation results in n_B bands of C1 feature maps. Each C1 band of feature maps is then convolved with to each of S2 patches with respected to their originally extracted locations. The set of S2 feature maps for an input image I (or **S2**_I) shall be obtained as follows $(b = 1 \dots n_B, i = 1 \dots N_f)$:

$$\mathbf{S2}_{\mathbf{I}} \equiv \{ (S2)_i^b = \eta \mathcal{M}_{E_{P_i}}[(C1)^b] * P_i \}$$
(1)

where $\mathcal{M}_A[B]$ is the masking operator of the feature map B located at local region of interest A; η is the normalization term to constrain the activation of the S2 map to [0, 1].

The final N_f -dimensional scale and position-invariant feature vector output is computed by taking the maximum operator across all bands and positions at C2 level. The C2 feature vector is used for scene analysis. The final C2 feature vector for an image I (or C2₁) shall be computed as follows $(i = 1 \dots N_f, b = 1 \dots n_B)$:

$$\mathbf{C2}_{\mathbf{I}} \equiv \{ (C2)_i = H_{\gamma}^+ [\max\{ (S2)_i^b] \}$$
(2)

Where:

max{A} is the maximum operator across all band b ∈ {1...n_B} and position (x, y) of the feature map A.
H⁺_γ[x] = max{x-γ,0}/(1-γ) (γ ∈ [0,1))

The list of parameters is given in the Appendix.

C. Feature Significance

It has been observed that the contribution of each feature in the gist vector towards the identification of a scene is different. Therefore, we propose a novel weighting scheme that characterizes the significance of each local feature in a gist vector towards final scene identification. Due to high level of redundancy in initial vocabulary construction, there might be a certain number of prototypes which is not informative, i.e. either ubiquitous or rare appearance, but could not be excluded initially due to no prior assumptions of the target environment. The significance measurement of each individual feature in our model is dependent on this activation profile throughout its temporal domain. The significance of uninformative features in the design should be reflected by low scores and vice versa. Hence, its effect is diminished for the final self-organization purpose. Based on this, the learning system can emphasize more on the salient group of features which mostly contributes to the scene categorization and also attenuate the effect of noisy features. The formulas for estimating significance of feature vector and its normalization are given as below $(i = 1 \dots N_f)$:

$$\begin{cases} S_{i} = \max\left(\frac{\max_{k \in \{1...t\}} \left(c_{i}(k)\right)}{\sum_{k=1}^{t} \left(c_{i}(k)\right)}, \frac{1-\min_{k \in \{1...t\}} \left(c_{i}(k)\right)}{\sum_{k=1}^{t} \left(1-c_{i}(k)\right)}\right)\\ \hat{S}_{i} = \frac{\Gamma_{\theta_{i}}(S_{i})}{\sum_{k=1}^{N_{f}} \Gamma_{\theta_{k}}(S_{k})} \end{cases}$$
(3)

where $\Gamma_{\tau}(x) = \begin{cases} x & \text{if } x \ge \tau \\ 0 & \text{otherwise} \end{cases}$ is the C2 feature vector from equation 3 at time step t. The equation 3 estimates the feature significance incrementally to the current time. It also introduces the competition between both the present and complementary absence part of feature activation[19]. In practice, the feature significance can be discovered bysupervised learning based on class labels as in [4]. However, unsupervised estimation based on feature correlations is critical for real-time navigation given situation such that supervision is not available or sometimes ambiguous in dynamic environment such as indoor. The threshold $\{\theta_i | i =$ $1 \dots N_f$ is used to filter low significance level and boost the contrast among features. In this model, all significance thresholds are set to $1/N_f$. Our mechanism suggests a principled way to select important features based on their temporal profile.

The concept of significance is not limited to only individual features but also to group of features which coactivate in spatial and temporal patterns. This is important in developing complex self-organizing patterns of features in cognitive neural network. The significance characterize features which may be selectively attended in different contexts [24]. Pertaining to the hippocampal episodic memories, the context-aware attention may be the triggered by the competition among various LTM cells to support anticipation and recognition of sequences.

IV. SPATIO-TEMPORAL SEQUENCE LEARNING

A. LTM Sequence Storage and Recognition

In this work, initial configurations of the target environment are clustered by the KFLANN (c.f. [5]). The KFLANN is an ART-based unsupervised network which offers fast learning of groups of scenes which have similar statistical correlations. The KFLANN architecture comprises of 2 layers: Input and Output. The input layer (F1) contains the input feature vector. The output layer (F2) contains output neurons which can be dynamically extended to accommodate new patterns. This fast learning is necessary for exploring data with no prior knowledge about number of clusters provided. The learning of the network is controlled by two two parameters: Vigilance ρ and Feature Tolerance set $\rho = \{\rho_i | i = 1...D\}$ where D is the dimension of input vector. The Tolerance Set ρ is determined by the standard deviation of the feature space as a means for controlling the feature uncertainties. The Vigilance parameter ($\rho \in (0, 1)$) characterizes the preferred generalization of the network. The activation function in vigilance testing of [5] is weighted by normalized feature significance. The weighting significance strategy ensures the clustering process attends to important group of features and also prevents the creation of noisy clusters.

The KFLANN in this case can be treated as vector quantization into visual tokens, which is similar to the concepts of character in text processing or a vector of Fourier transforms in speech recognition. The number of KFLANN iterations to stabilize the centroids by reshuffling the data and is empirically set to 5. Each visual token corresponds to a set of locations which share similar C2 feature properties. When a new input image arrives, its extracted C2 feature vector is presented to the KFLANN network. If this feature vector satisfies the vigilance testing in equation 4, the winning neuron C_w in F2 fires with the strength F_{C_w} as in equation 6. This firing will then update the state for the connected LTM cells as shown in Figure 1. If the existing LTM cells fire below the recognition threshold, a learning signal is triggered and new output neuron is extended to accommodate this new pattern sequence. The KFLANN algorithm is presented as in Algorithm 1.

B. LTM Sequence Storage and Recognition

We adopt the same terminology as in [20], [21] in this model. A temporal sequence S is defined as $S: S_1 - S_2 \dots - S_N$ where N and $S_i(i = 1 \dots N)$ is the length and a component of the sequence respectively. Any $S_i, S_{i+1} \dots S_j$ where $1 \le i \le j \le N$ is called a subsequence. If S contains repetitions of the same subsequence, it's called a complex sequence, otherwise a simple sequence.

In our sequential memory model, each LTM cell is dedicated to a sequence. One representation of sequence by a

Algorithm 1 KFLANN Clustering

Notations:

- $C2_I$: Input vector of D features $C2_I = \{c_i | i = 1 \dots N_f\}$.
- w_{ii} : Synaptic weight from input feature *i* to output node *j*.
- J: The current number of active (committed) output nodes.
- C: The temporary output candidate list of each input $C2_I$.
- ρ : Vigilance parameter
- $\rho_i (i = 1 \dots N_f)$: Tolerance of input feature vector.

Begin Algorithm:

- $\mathcal{C} \leftarrow \{\emptyset\}$
- $J \leftarrow \dot{0}$
- $\delta_i \leftarrow StdDev(c_i)$
- for each $C2_I$ at F_1 do
 - for each $j \in F_2$ do

Calculate the matching function:

$$T_{j} = \sum_{i=1}^{N_{f}} \left\{ \hat{S}_{i} \cdot \mathbf{1} \left\{ \delta_{i}^{2} - ||w_{ji} - c_{i}||_{2}^{2} \right\} \right\}$$
(4)
where $\mathbf{1}\{a\} = \left\{ \begin{array}{cc} 1 & \text{if } a > 0 \\ 0 & \text{otherwise} \end{array} \right.$
if $\{T_{j} \ge \rho\}$ then
 $\mathcal{C} \leftarrow (\mathcal{C} \bigcup j) \$ {Vigilance Test succeeds}
end if
end for
end for
if $\mathcal{C} \equiv \emptyset$ then
Create new $(J+1)^{th}$ F2 node.
Direct mapping from $C2_{I}$ to the weight of the node $(J+1)^{th}$.
 $J \leftarrow (J+1)$

else

if

for each $j \in \mathcal{C}$ do

Calculate the neural activation:

$$F_j = ||W_j - C2_I||_2^2 \tag{5}$$

where $W_j = \{w_{ji} | i = 1 \dots N_f\}$

end for

Determine the winner centroid by Winner-Take-All rule:

$$\mathcal{C}_w = \operatorname{argmin}_{i \in \mathcal{C}} \{F_j\} \tag{6}$$

Assign index of $C2_I$ to centroid \mathcal{C}_w

end if

Recalculate centroid coordinates by mean of its members. Reshuffle all centroid's 1-nearest-neighbor to top of the data set Reset output node F2 and weight vectors.

group of LTM cells is discussed in [7]. The output of each LTM indicates the similarity between its stored sequence and the input sequence. The structure of a LTM is shown in Figure 2. It comprises of consecutive pairs of Primary neuron (PN) and Dual Neuron (DN) [21]. Each PN neuron receives the feedforward excitations from the STM output neurons. The sequence length in each LTM cell is determined by the number of PN/DN pairs. Each DN serves as internal STM for each LTM cell to update the next element of the sequence tracking. During storage phase, each sequence is stored in the corresponding LTM cell via a one-shot learning mechanism. Since the input sequence may be complex, there may be multiple connections from each STM neuron to PNs. The recognition algorithm in each LTM is given in Algorithm 2.



Fig. 2. A LTM Cell Structure

When a new image is presented at the input, all PN neurons are updated concurrently (Equation 8) according to the input vector. Subsequently, DNs are updated sequentially (Equation 10) and the best similarity score is propagated to the last DN of the LTM (Equation 11). The delay τ controls the amount of tolerated latency in signal arrival. The algorithm is designed such that the original learned sequence will elicit highest response N, i.e. perfect match. Deviations from the original sequence will lower the matching score proportionally [6]. The complexity for PN/DN updating operation in Algorithm 2 for each LTM cell is approximately O(N) where N the length of the cell. During retrieval phase, all LTM cells compete based on their similarity outputs and the best matching cell is declared as the winner. The recognition algorithm is able to play sequences continuously without specifying starting and ending points.

V. EXPERIMENTS

In our experiments, we used the Robot Vision dataset from the ImageCLEF Competition 2010 [25]. The images were captured sequentially from a mobile robot which moved around different locations in the same building. There are 9 different places within the same building. The locations of place categories are: Corridor (C), Elevator (E), Kitchen (K), Lab (L), Large Office 1 (LO1), Large Office 2 (LO2), Printer Area (PA), Small Office (SmO) and Student Office (StO). One category may contain several sequences since each place may be visited multiple times. Additionally, each position of capture contains images from both the left and right camera mounted on top of a mobile robot. There is only a small displacement between the viewpoints of the left and right camera. We denote the set of images of the left camera Set S and that of the right one Set R which contains moderate pose difference from Set S. We used the image sequences from Set S for sequences storage and both the Set S and the Set R for sequence retrieval evaluation.

For feature building and extraction, we randomly extracted 10% of the images in each category of Set S to construct our S2 prototypes. Subsequently, 2000 patches were randomly selected for feature extraction. The feature significance is

Algorithm 2 LTM Sequence Recognition

Notations:

- I_i : Input of the PN i.
- PN_i : Excitation value of the PN i.
- DN_i : Excitation value τ of the DN *i*.
- τ : Delay factor.
- Del_i : Delay counter of the DN *i*.
- O: Output activation
- N: The length of the sequence.
- W: The LTM weight matrix.

Begin Algorithm:

- $DN_i \leftarrow 0, i = 0 \dots N$
- $Del_i \leftarrow \tau, i = 1 \dots N$

For each presented image, the C2 feature vector is extracted and fetched into KFLANN network. The input excitations to LTM cells are calculated as $(i = 1 \dots N)$:

$$I_i = \begin{cases} (1 - F_i) & \text{if } T_i \ge \rho \\ 0 & \text{otherwise} \end{cases}$$
(7)

where F_i follows Equation 7 if vigilance testing succeeds. Update all PN neurons:

$$\mathbf{PN} = \mathbf{W} * \mathbf{I} + \mathbf{DN}_{-1} \tag{8}$$

where: $\mathbf{PN} = \{PN_i | i = 1...N\}$, $\mathbf{I} = \{I_i | i = 1...N\}$ and $\mathbf{DN}_{-1} = \{DN_{i-1} | i = 1...N\}$.

for i = 1 to N do if $\{(Del_i \ge 0) \land (DN_i \ge \max\{DN_{i-1}, PN_i\})\}$ then

$$Del_i \leftarrow Del_i - 1$$
 (9)

else

$$\begin{cases} Del_i = \tau \\ DN_i = \max\{DN_{i-1}, PN_i\} \end{cases}$$
(10)

end if

end for

Update the output neuron activation:

$$O = DN_N \tag{11}$$



Fig. 3. Sample Images from the ImageCLEF2010 Dataset

then estimated based on the same set of images. We constructed our visual tokens by using the KFLANN clustering for Set S with Vigilance $\rho = 0.7$. A total number of 81 clusters were collected with this vigilance. The clustering stage to form STM clusters is not purposefully tailored for any particular category. This is to leverage the roles of sequential property in recognizing places. To establish each sequence, each input image was mapped to winning STM cell as in equation 6. A small number of images which did not satisfy any vigilance testing was rejected and did not participate in the sequence construction.

As previously mentioned, sequences are stored in LTM

cells via one-shot learning. For sequence retrieval, the exact sequence as in the LTM cell will elicit highest response which has the magnitude of the length of the cell. Therefore, if each full sequence is stored in a single dedicated LTM cell, its length directly biases the winner decision by the WTA competition over all LTM cells. Thus, in our experiments, the longer sequences are broken into consecutive LTM cells of similar length $N_{LTM} = 100$ Sequences which are of shorter length than N_{LTM} are concatenated with the end part of the previous sequences so as to make the equal length. One of the possible solutions to the length problem was discussed in [21] where the sequences can be chunked and stored in a hierarchical fashion. The number of stored LTM cells in each category was listed in Table 1. The winning LTM is decided by the most excited LTM cell, and the location is determined by its corresponding location identifier.

During testing, we played the sequences continuously in the order as in the last row of Table 1 according to the original capturing trajectory (4040 images). However, any arrangement of sequences is possible. The dynamics of all the LTM cells during retrieval of original Set S are depicted in Figure 4. While navigating, each LTM cell competes with others and the excited LTM cells activation will gradually increase to its highest possible potential and then decrease when the robot moves in and out of its coverage. Smooth transition among LTM cells may be obtained if certain degree of overlapping in consecutive cells is imposed. The design of temporal overlapping is not within the scope of this paper but can be implemented as in [7]. At each time, not only the information about the location can be obtained from the winning cell, the level of confidence of being in certain location can be derived from the strength of the winner. By analyzing the dynamics in LTM cells activation further anticipation can be made. At each location, we can see there are clear separations of activations between its LTM cells and other locations. Therefore, by exploiting the sequential property the system can localize itself even though the universal set of elements is shared among sequences. The accuracy estimation of a category is defined by the number of matches between the ground truth and the label of the maximum response LTM. This estimation is also called oneshot classification [12]. The accuracy of sequence retrieval with Set S and Set R is shown in Table II.

The strength of the winning cells response at each time is proportional to the predictions confidence of one single data to be at a particular location based on its entirely previous history. However, if its magnitude is low, location prediction is unreliable despite being the winner. This can be illustrated by when one moves from one location (e.g. kitchen) to another location (e.g. corridor), the place between the two locations should not be classified as solely one or the others but the transition between two places (Figure 4 (Upper)). This property can be observed by the dynamics of our sequence learning framework by considering the whole episodes from one place to another place. Figure 4 (Lower) illustrates the LTM activation profile when the robot moves

 TABLE I

 Sequences properties with different location categories

Category	С]]	Ξ	К	L	LO1	LO2	PA	SmO	StO			
Sequence Index	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	E_1	E_2	K_1	L_1	$LO1_1$	$LO2_1$	PA_1	SmO	StO_1
Length	228	94	134	33	32	232	164	71	133	57	32	29	599	492	351	508	170	355	326
Number of LTMs	3	1	2	1	1	3	2	1	2	1	1	1	6	5	4	6	2	4	4
Playing Order	Playing Order $C_1 - K_1 - C_2 - C_3 - C_4 - LO1_1 - C_5 - LO2_1 - C_6 - SmO - C7 - StO - C_8 - L_1 - C_9 - PA_1 - C_{10} - E_{11} - E_{12}$																		

TABLE II ACCURARY (%) OF LOCATION RECOGNITION BASED ON INDIVIDUAL IMAGE WITHOUT THRESHOLDNG

Category	С	Е	К	L	LO1	LO2	PA	SmO	StO
Set S	84.9	96.7	84.8	91.7	61.3	74.2	68.3	86.8	85.6
Set R	78.6	78.7	85.6	84.8	73.2	85.4	65.3	76.3	81.3

TABLE III

Accurary (%) of location recognition based on individual Image with thresholdng

Category	С	Е	K	L	LO1	LO2	PA	\mathbf{SmO}	StO
Set S	96.4	97.0	93.6	99.7	90.3	93.7	99.0	94.2	99.6
Set R	80.3	72.3	88.3	98.0	87.4	100.0	88.8	87.5	94.6

TABLE IV

ACCURACY AND STANDARD DEVIATIONS OVER 10 TRIALS (%) FOR MISSING ELEMENTS SEQUENCES

p (%)	100	90	80	70	60
Set S's Accuracy	84.03	83.81	83.25	79.03	78.06
Set S's Std Dev	0.0	0.6	0.4	0.5	0.5
Set R's Accuracy	80.25	79.48	78.35.25	76.72	74.73
Set R's Std Dev	0.0	0.5	0.4	0.3	0.4

We sampled randomly p(%) images of each sequence of Set R. However, the internal temporal ordering in each sequence was still preserved. The alteration degrades the maximum possible matching score of shortened sequences proportionally comparing to the full-length sequences. This may be the case when certain data are missing under various situations while capturing. We played them in the same order as the last row in Table I. We conducted 10 trials for each value of p without thresholding by θ for comparison. The average accuray over all categories for the two sets is reported in Table IV. The accuracy when p is 100% (full length) is derived from the average of all categories reported from Table II. We can see that the accuracy gradually drops as the length of each sequence is reduced. However, the standard deviations over 10 trials are small ($\leq 1.0\%$). The result justifies the stability of our sequence recognition under length distortions by exploiting the sequential property.

VI. CONCLUSIONS

This paper presents a novel hierarchical architecture based on the interaction between STM and LTM mechanisms for spatio-temporal sequence learning. We explained the generic feature building and extraction, STM and sequence storage and recognition by the LTM organization. We also analyzed the efficacy of the proposed framework in a visual localization application. We showed that our system is able to localize continuously based on competition of LTM cells. In the experiments, we used the universal set of elements to construct the sequence of many different locations to substantiate the power of sequential property. Additionally, the stability of our sequence learning architecture was also

from the room Kitchen to the Corridor. The transition state between two environments can be observed by the gradual decrease in activation of the Kitchen LTM and the increase in Corridor LTM. By this, the decision of place should be made only when the activation of the LTM is sufficiently large. It can be implemented by imposing a threshold θ for decision making only when the winner activation exceeds θ .

Table III shows the result of the Location recognition by the sequence retrieval with $\theta = 0.4$. This sequence refused to classify for approximately 10% of the images for Set S and Set R. However, it significantly improves the accuracy when decisions were made. The determination of the threshold θ depends on the practical tolerance acceptance. For instance, it can be estimated based on the average activation of the LTM cell to random noises of the sequence. The automatic estimation of this threshold will be subject of future work.



Fig. 4. Upper - Example of ambiguity of place. The left two images are labeled Kitchen while the right two images are labeled Corridor from the dataset. Lower - Activation of the last LTM of category K and the next LTM of category C

To illustrate the robustness of the sequence retrieval against the variations in the length of testing sequence.



Fig. 5. Different LTM activations of each category during recall rate using Set S. The playing order follows the last row of Table 1

demonstrated with certain distortions, i.e. length variations. Further extensive evaluations on the capability of the systems towards other robustnesss evaluation such as distortions of the robots trajectory are not within the scope of this paper.

Our intention is to integrate this architecture to build a hierarchical episodic memory model which characterizes various interactions and self-organizations between STM and LTM mechanisms. The architecture can be extended to facilitate many components of embodied intelligence including sensory input processing, anticipation, motor control and goal creation with robust tolerance[26].

Label	Layer	Values
$n_{\mathcal{S}}$	S1	16
$n_{\mathcal{O}}$	S1, C1	4
$N_s^{\mathcal{G}}$	S1	$\{(2n+1) n=3,\ldots,18\}$
$N_o^{\mathcal{G}}$	S1	$\{0,\pi/4,\pi/2,3\pi/4\}$
$n_{\mathcal{B}}$	C1	8
$n_{\mathcal{GS}}$	C1	8
$N_{i\in\{1,\ldots,n_{GS}\}}^{GS}$	C1	$\{(2n) n=4,\ldots,11\}$
$n_{\mathcal{P}}$	S2	4
$N_{i\in\{1,\ldots,n_{\mathcal{P}}\}}^{\mathcal{P}}$	S2	$\{10, 20, 30, 40\}$
γ	C2	0.6

APPENDIX

REFERENCES

- [1] M. Bar, "Visual objects in context," Nat. Rev. Neuroscience, 2004.
- [2] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *Journal of Physiology*, vol. 195, pp. 215–243, 1968.
- [3] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Bio. Cybs. - Springer Verlag*, vol. 36, pp. 193–202, 1980.
- [4] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *PNAS*, 2007.
- [5] A. Tay, J. Zurada, L.Wong, and J. Xu, "The hierarchical fast learning artificial neural network (hieflann)-an autonomous platform for hierarchical neural network construction," *IEEE Trans. Neural Network*, vol. 18, no. 6, pp. 1645–1657, 2007.
- [6] J. A. Starzyk and H. He, "Spatio-temporal memories for machine learning: A long-term memory organization," *IEEE Trans. Neural Network*, vol. 20, no. 5, May 2009.

- [7] J. A. Starzyk and H. He, "Anticipation-based temporal sequences learning in hierarchical structure," *IEEE Trans. Neural Network*, vol. 18, no. 2, March 2007.
- [8] J. O'Keefe, "Place units in the hippocampus of the freely moving rats," *Experimental Neurology*, vol. 51, pp. 78–109, 1976.
- [9] A. D. Redish, Beyond the Cognitive Map: From Place Cells to Episodic Memory. MIT Press, 1999.
- [10] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelop," *IJCV*, vol. 42, no. 3, pp. 145– 175, 2001.
- [11] L. W. Renninger and J. Malik, "When is scene identification just texture recognition," *Vision Research*, 2004.
- [12] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. PAMI*, vol. 29, no. 2, pp. 300–312, 2007.
- [13] S. Ullman, "Visual features of complexity and their use in classification," *Nat. neuroscience*, vol. 5, no. 7, pp. 682–687, 2002.
- [14] R. Epstein and N. Kanwisher, "A cortical representation of the local visual environment," *Nature*, no. 6676, pp. 598–600, 1998.
- [15] K. Tanaka, "Inferotemporal cortex and object vision," Ann. Rev. Neuroscience, vol. 19, pp. 109–139, 1996.
- [16] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. Howard, H. W., and L. D. Jackelm, "Backpropagration applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [17] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat. Neuroscience*, vol. 2, no. 11, 1999.
- [18] H. Eichenbaum, P. Dudchenko, E. Wood, M. Shapiro, and H. Tanila, "The hippocampus, memory and place cells: Is it spatial memory or a memory space?" *Neuron*, vol. 23, pp. 209–226, 1999.
- [19] G. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer Vision* and Image Understanding, vol. 37, pp. 54–115, 1987.
- [20] D. Wang and M. Arbib, "Complex temporal sequence learning based on short-term memory," *Proceedings to IEEE*, vol. 78, no. 9, pp. 1536– 1543, 1990.
- [21] D. Wang and M. arbib, "Timing and chunking in processing temporal order," *IEEE Trans. SMC*, vol. 23, no. 4, pp. 993–1009, 1993.
- [22] D. Wang and B. Yuwono, "Anticipation-based temporal pattern generation," *IEEE Trans. SMC*, vol. 25, no. 4, pp. 615–628, 1995.
- [23] N. Burgess and G. Hitch, "Computational models of working memory: putting long-term memory into context," *Trends in Cognitive Sciences*, vol. 9, no. 11, pp. 535–541, 2005.
- [24] A. V. Samsonovich and G. A. Ascoli, "A simple neural network model of the hippocampus suiggesting its pathfinding role in episodic memory retrieval," *Learning Memory*, vol. 12, pp. 193–208, 2005.
- [25] ImageCLEF, "http://www.imageclef.org/2010/icpr/robotvision," 2010.
- [26] J. A. Starzyk, *Motivation in embodied intelligence*, ser. Robotics , Automation and Control. I-Tech Education and Publishing, Vienna, Austria, 2008.