Robotic and document analysis cross-fertilization: Improving place cells based robot navigation

Dalia Marcela Rojas Castro, Arnaud Revel, Michel Ménard

Robotic and Document Analysis Cross-Fertilization: Improving Place Cells Based Robot Navigation

Dalia Marcela Rojas-Castro, Arnaud Revel, Michel Ménard
Computer Science Department, Laboratory L3i of La Rochelle University
Avenue Michel Crépeau 17042, La Rochelle, France
{dalia_marcela.rojas_castro, arnaud.revel, michel.ménard} @ univ-lr.fr

***Abstract*—This paper proposes a place cell model allowing place recognition in the context of robot autonomous navigation. The robustness of this approach lies in the fact that even if one or several patterns characterizing the place are removed or not visible anymore, a place can still be recognized. The recognition process in this work is improved with respect to the state-of-the-art place cells approach. Additionally, the interconnection of the modules is made such that the robot is able to learn new places as it navigates and interacts with the environment to get to its final destination. Experimental results validate the advantage of the incremental learning allowing the robot to cope with any unforeseen changes and thus adapting itself to the environment.**

Place-cells recognition; robot navigation; visual perception; data merging, hybrid control architecture.

I. INTRODUCTION

Robot navigation based on visual perception systems (such as onboard camera systems) has been especially prevalent over the last three decades. These systems are robust and reliable as they provide detailed information about the environment, which may be overlooked by other types of sensors.

Several researches have focus their work on visual place recognition, particularly in the context of autonomous robot navigation, where different approaches have been proposed for solving mapping and robot localization problems. They endow the robot the capacity of "understanding" its surrounding environment, knowing its position with respect to a reference point and thus facilitating its navigation task.

Place recognition is used since a place can be identified as a stable reference point that can be learned by keeping in memory the location of the most relevant perceived patterns within the panoramic visual field of the robot. Returning to this place then consists in navigating until recognizing the same learned patterns.

In the state-of-the-art, Gaussier [1] has proposed a biologically inspired approach for recognizing places within the navigation environment. The robustness of this approach lies in the fact that even if one or several patterns characterizing the place are removed or not visibly available anymore, a place can still be recognized. Additionally, by means of a triangulation process it is possible to obtain information about the robot's position with respect to the surrounding environment. Therefore, the underlying idea of this paper is to extend this system to allow a robust navigation of a humanoid robot...

Research work supported by European regional development Funds(Contract35053) and the Poitou Charente Region

Our approach addresses the problem in a manner that differs mainly in two points:

Firstly, the procedure for detecting the landmarks undertakes two-classification process. The SIFT local descriptor [2] and a visual bag of words model are first used in order to describe distinctly the salient features of all the images. Then, the features are clustered according to their proximity in terms of distance and the resulting groups are considered as salient landmarks. Finally, each landmark is compared to others by computing the norm of the difference between the features describing them.

Secondly, the internal computation of the neural components are modified in order to allow the robot to compare the landmarks perceived from different places during navigation, by using a vigilance term inspired by the work of Grossberg [3] and learn them when not recognized. Consequently, the system learns incrementally.

This system has been conceived such that it is able to interact with the other layers of our control neural architecture that we have previously presented in [4] and [5]. The overall architecture is based on a combination of a top-down and bottom-up approach. It allows the robot to emulate human behavior in navigating inside an unknown building by "reading" a map or a floor plan with its camera and "remembering" a sequence of signs to follow on the way.

In this article, by injecting the place recognition model into our previous architecture, we make the robot capable of coping with any unforeseen changes such as the occlusion of the expected signs. It allows the robot to find new references points by itself and learn them to get to its final destination.

II. RELATED WORK

The visual place recognition problem is usually tackled by retrieving images of the scene, training them and comparing them to other images thereafter. The undertaken process for place recognition can be considered as a more generalized version of the pattern recognition process and can be distinguished in two types of recognition depending on the application: topological place recognition and place categorization. In the context of robot navigation, a topological place recognition consists in endowing the robot with the capability of recognizing previously observed places in known environments. Appearance-based SLAM techniques such as the FAB-MAP [6],[7],[8] convert the images from a set of local features into a bag-of-words representation in order to match the appearance of the current scene to train the data. Place

categorization instead, allows the robot to give a semantic labeling to the places by classifying different locations of a new environment into categories [9], [10], [11].

Biological approaches propose place recognition models which emulate similar behaviors seen in living organisms based on allothetic information when performing goal-orientated navigation tasks [12], [13]. Indeed, many studies on insects like bees, ants and wasps have shown that they could store multiples views of a place with different positions in order to learn the place. Then, by comparing and matching the stored images to the newly perceived one, they are able to recognize the place and return to their nests [14], [15]. Similarly, neurobiological studies in mammals like primates and rats have revealed that they also use surrounding visual cues in order to achieve a particular place in a comparable way to insects. However, mammals show higher generalization capabilities and more complex processing when performing recognition of a scene or a place. Experiments conducted by 0'Keefe & Dostrovsky [16] led to the discovery of pyramidal neurons in the rat hippocampus that fire at their maximal activity when the animal is at a particular location in the environment and decreases as it goes away from it. They are called "place cells" and the regions at which they fire at are called place fields, which are almost similar to the receptive fields of sensory neurons. Based on this insight, Gaussier [17] and Giovannangeli [18], proposed a model of the hippocampal systems where "place cells" are learned as a result of the recognition of a particular configuration merging which our model is based on.

The robustness of such model lies in the use of the spatial information of the set of local features composing the place in addition to the description of the place itself as seen in some of the SLAM techniques.

III. Overview of the Hybrid Neural Control Architecture

Before entering into the core of the Place-cells recognition system, an overview of the overall architecture is first given. The complete architecture integrates a floor plan analysis into an organized neural structure. It is composed of three modules (as illustrated in Fig.1): a deliberative module and two reactive modules. The deliberative module corresponds to the processing chain in charge of extracting a sign sequence from the floor plan by computing a navigation path. On one hand, the analysis of the floor plan in real time, represented by the box in the top half of Fig.1, undertakes a thorough process permitting the robot to extract the relevant information for its integration into the system. It consists of (1) an information segmentation process, which identifies and separates different types of information; (2) followed by a structural analysis where the information is extracted (walls and navigation signs separately); (3) and finally a semantic analysis allowing the extraction of the sign sequence based on the computation of the path and the information of the signs.

On the other hand, the neural structure of the two reactive modules represented by the two boxes in the lower half of Fig.1 follows a perception-action mechanism that constantly evolves because of the dynamic interaction between the robot and its environment. Whereas the first reactive module integrates the sequence information coming from the deliberative module and constantly uses it in order to control its online navigation [5]; the second module learns places as new reference points when the expected signs are not visibly available. These modules are further composed of two levels of data streams corresponding to perception and action flows. The first level uses a reflex mechanism that controls directly the robot's action based on

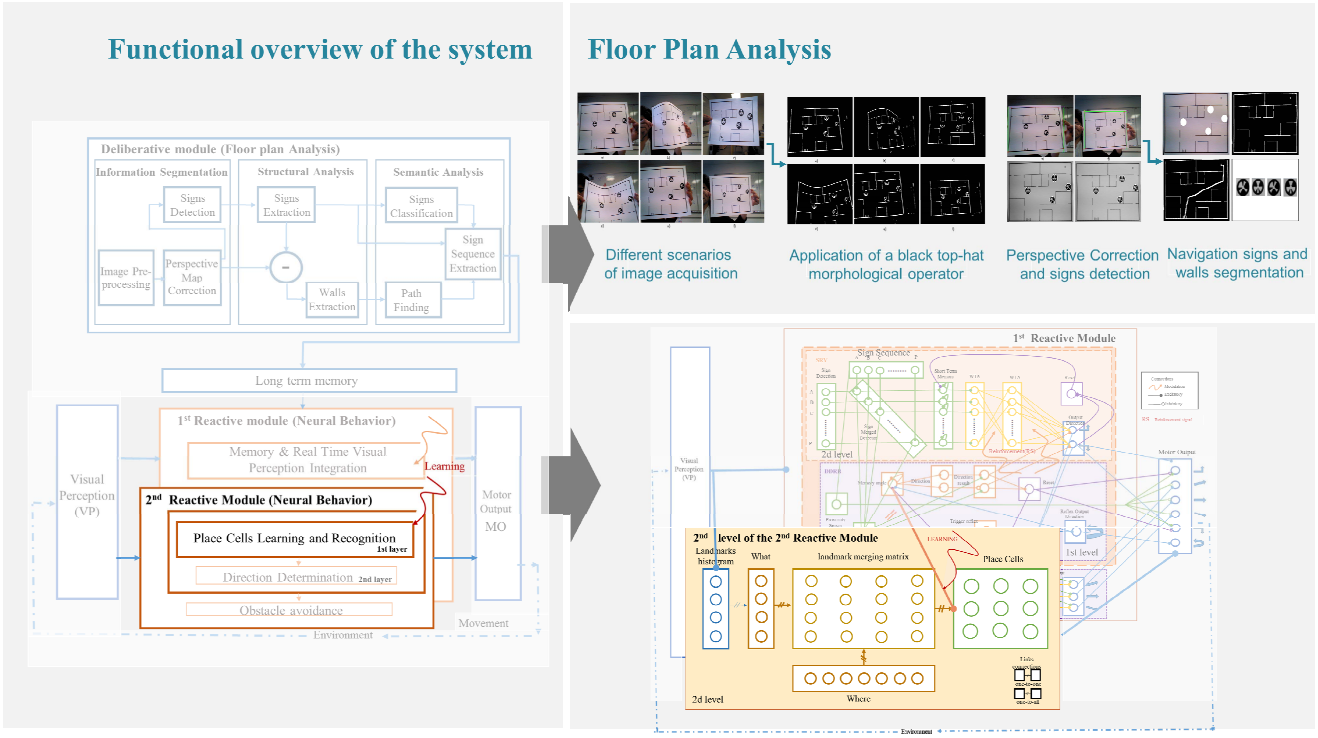

Figure 1. Functional overview and implementation of the neural architecture

the information extracted from the perceived input. The second level uses a cognitive mechanism performing recognition of the perceptive flow.

Since the whole system works in parallel, a competitive mechanism allows to decide on the best behavior for controlling the robot according to the stimulus received. Hence, the neural interconnection is done by either excitatory or inhibitory connections allowing or preventing the activation of neurons respectively.

IV. PLACE CELLS LEARNING AND RECOGNITION LAYER

A. *Natural Landmarks detection and extraction*

A robust visual place recognition algorithm needs to combine descriptive, discriminative and generalization properties. Therefore, in order to capture all these properties this paper presents a model for place recognition, which learns and recognizes places. Each place is represented by place cells resulting from the recognition of a particular configuration merging two flows of information: *what* and *where* [19]. While, the former is in charge of identifying and recognizing the patterns (landmarks in this work) perceived in the visual scene of the place, the latter is involved in the analysis of the spatial location of the same. In order to achieve this, it is necessary to, on one hand, describe the perceived landmarks in a distinctive way and on the other hand, find their respective location within the scene (section A details the extraction of such information).

Figure 2 illustrates the complete neural layer allowing to learn different places (9 places tested in this work as shown in the right part of the figure) of a given environment. Each place is characterized by a set of landmarks surrounding the robot within its panoramic view (360°) and the information extracted from all landmarks is used as the input of the network system: On one side, the information related to the description of each landmark is gathered in the *landmark histogram neural group*, which is connected, to the *what group*. On the other side, the information related to the position of each landmark, given by its X-coordinate within the image with respect to a global reference point, is immediately fed into the *where group* to be processed. Then, the information coming from both groups converges on a two-dimensional array of neurons, which keeps in memory the resulting value of all the landmarks perceived in the panorama. As a result, a landmark constellation is formed in the *landmark merging matrix group* leading to learn a new place by recruiting a new neuron in the *place cell group*.

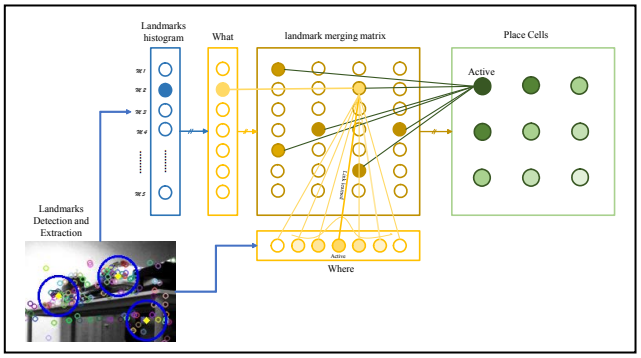

Figure 2. Place cells learning and recognition layer. A place cell is formed by the combination of two neural groups (*what* and *where*) carrying the information from the extracted landmarks.

B. *Natural Landmarks detection and extraction*

In this work, the natural landmarks are considered as patterns that can be described by their distinguishable features (key-points). Hence, in order to detect these patterns all images acquired by the sensor camera are first pre-processed and transformed into gray color space. Then, for each image, the salient features along with their descriptors are computed. Since the viewpoint of a pattern can drastically change from one position to another, the SIFT descriptor is used. As a result, each feature is described as unique and differ from one another according to the composition of their visual characteristics.

This process is performed for two different sets of images: A set of template images to build the vocabulary and a set of query images to build, detect and recognize the landmarks during the robot run-time navigation.

1) *Vocabulary construction*

The set of template images used here depicts the navigation environment from different viewpoints.

For each template image, all interest features (key-points) are detected and their descriptors are computed with the SIFT descriptor and stored in a vector. Then the k-means algorithm is used in order to group all the descriptors into k different clusters according to their similarity.

This whole process is only performed once; therefore, it was computed before the robot navigation. At the end, the clusters centers centroids are stored and used as inputs in the following stage.

2) *Landmark construction*

In order to build the landmarks, the extracted features from the query images are first compared and assigned to the clusters of the vocabulary according to the descriptors similarity. Then, the same descriptors are grouped into k different clusters according to their spatial distance (see Fig. 3a). This is done by using the Euclidean distance and the x and y coordinates of each key-point. The k number of clusters corresponds to the total number of landmarks in each image composing the panoramic view.

As a result, each spatial cluster is considered as a "natural landmark", which comprises of a certain number of key-points close to each other and belonging, each of them, to different clusters of the vocabulary.

3) *Landmark representation*

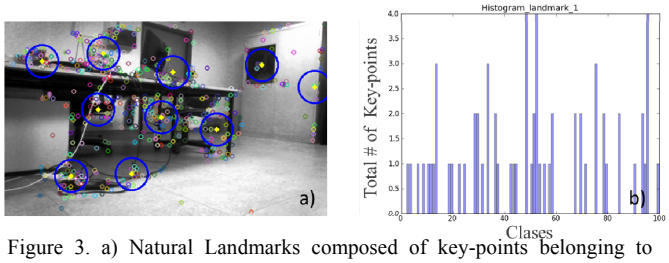

Figure 3. a) Natural Landmarks composed of key-points belonging to different classes according to their descriptors similarity. b) Landmark histogram of size 100.

Since the number of key-points belonging to each "natural landmark" varies from one another, it is not possible to compare them by the same standards. Therefore, each landmark is represented by a histogram composed of as many bins as there are clusters (size *n*) in the vocabulary. Then, for each landmark, each bin counts the total number of key-points that are associated with the cluster corresponding to that bin.

As a result, all landmarks have the same number of parameters to be compared to, regardless the number of key-points comprising them (see Fig.3b). Henceforth, each landmark is characterized by three elements:

- A Landmark histogram based on a *n*-sized vector
- An x-coordinate, corresponding to the relative position of the landmark-cluster center with respect to the horizontal axis of the image ;
- A y-coordinate, corresponding to the relative position of the landmark-cluster center with respect to the vertical axis of the image.

C. *Landmark histogram group*

This group comprises of as many neurons as there are feature classes in the system, which corresponds to the vocabulary clusters. Thus, each neuron corresponds to one type of feature class and their activity value is computed by calculating the number of key-points comprised within the landmark belonging to each class. In other words, the neurons are considered as bins, where each key-point belonging to a landmark is thrown into its corresponding neuron class. Thus, a form of histogram is created for each landmark where each neuron of *the landmark histogram group* adds up the number of key-points that belong to it as shown in figure 4. If for instance, there is no key-point belonging to a given class inside the landmark, the final value of its corresponding neuron class is zero.

Once the activities of all neurons have been computed, the resulting vector is injected into the *what* group to be either learned or recognized. Afterwards, the activity values of the neurons are reset to zero so that the same process can be repeated with the other landmarks of the same panorama view.

D. *What group*

This group is composed of a sufficient number of neurons to encode the total amount of landmarks that can be found in a given exploration environment. Two different procedures have been implemented, one for learning and the other for recognizing. The first one consists in recruiting and learning one neuron for every landmark perceived in the panoramic view of the first place. The recruited neuron is then associated with the landmark by performing a one-shot learning. The synaptic connection weights are modified according to the following rule.

$$\Delta W_{ij} = a_i * a_j \quad (1)$$

Where, the current activity neuron a_j equals to 1 when the *what* neuron is recruited and 0 otherwise; and a_i is the activity of the landmark input neuron.

The second procedure consists in comparing the landmarks of a new place to those already learned to find out if the new landmarks can be recognized. To this end, all the neurons activities are computed accordingly to the following equation:

$$a_j = 1 - \left(\frac{\sum_{i=0}^{L_i} | W_{ij} - a_i |}{L_i} \right) \quad (2)$$

Where, a_j gets the maximum value when the Euclidean distance between the synaptic weights W_{ij} and the input value a_j are nil.

The resulting values determine how similar the current landmark is to the previously learned. Then, by verifying if its activity value is bigger than a "vigilance" term, it is possible to conclude that the landmark has been recognized. Otherwise, it has to be learned as a new one by following (1).

E. *Where group*

The *where* group, is composed of a limited amount of neurons encoding a preferred direction covering in all the total 360 degrees of the panoramic view in order to compute the position of each landmark. As the angular position of each landmark is given by its distance in pixels with respect to the reference point (azimuth), the preferred directions are also given in pixels. For every landmark position, the *where* group is expressed as a non-normalized Gaussian activity profile:

$$a_j = exp - \frac{(\alpha - \mu_j)^2}{2\sigma^2} \quad (3)$$

Where, α represents the azimuth of the j_{th} landmark and μ_j the preferred direction of the neuron *j*. Each preferred direction is computed as:

$$\mu_j = \frac{dist}{2} + k * dim \ , k\ (1,2,3 \ldots dim) \quad (4)$$

Where, *dist* is the distance in pixels between each preferred direction and *dist*= X_{total} / *dim*. With, X_{total} the total number of pixels in the panorama view and *dim* the number of neurons in the *where* group.

The same computation is performed for each neuron in the group with the same α. Since their preferred direction is different, only the neuron with the closest value to α results with the maximum activity value, and consequently gets to

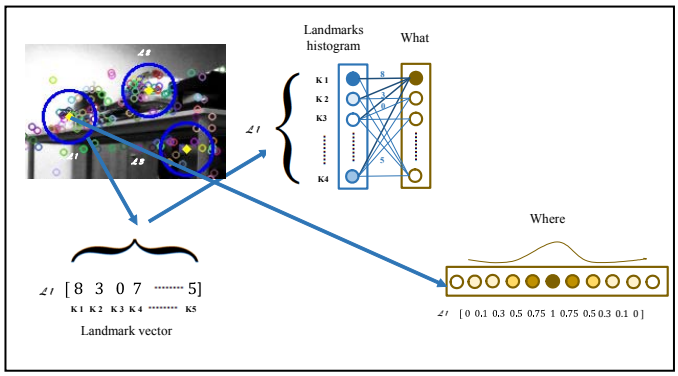

Figure 4. Computation of the what and where groups

encode the landmark position.

Additionally, by means of a lateral diffusion, its neighbor neurons are set to a value that decays in function of their distance among them in order to allow a better tolerance in the robustness.

F. Spatio-temporal merging matrix and place cells learning

In order to learn a place, the robot needs to keep in mind the information of all the surrounding landmarks perceived from its point of view. However, as the analysis of the place, which is given by the analysis of the landmarks within the panoramic view, can only be done in a sequential mode (the system cannot recognized several landmarks in parallel), it is necessary to keep in memory the overall information.

Thus, in order to suppress the sequential aspect of the scene exploration, a matrix of neurons stores the information of all landmarks perceived in the panoramic view. In fact, the information coming from both *what* and *where* groups of each landmark converges into the matrix allowing a spatio-temporal integration. The activity of the detected landmark in the *what* group and the activity of its position in the *where* group, enable the activity of the corresponding neuron in the matrix merging group. As a result, after having processed all the landmarks from a place, a landmark constellation is formed allowing to learn a new location by encoding a neuron in the Place cell group (see Fig.2).

V. EXPERIMENTAL RESULTS

A. Procedure

In order to validate the proposed approach, the place cell learning and recognition layer was implemented in the humanoid robot NAO and tested in a 2.5m x 2.5m room space.

To this end, a simple scenario was created such that a certain number of positions scattered all over the test environment was chosen to code different places within it in order to allow the robot to depict its environment by learning and recognizing each place at which it is located.

Figure 5 (top part) depicts the test environment composed of nine different places representing nine different positions covering in all the entire navigation test environment.

Each place is characterized by the set of patterns (landmarks) and their corresponding positions perceived by the robot within a panoramic view (360°). The position of the landmarks is calculated with respect to a north (here an artificial landmark was used as the global reference point as illustrated in the top part of figure 5).The panoramic view is obtained by taking snapshots of the environment with the camera's robot while rotating around itself. In overall, twelve overlapping images were enough to complete one panoramic view and thus build a robust representation of a place. As illustrated, each image is composed of ten landmarks, but as each image overlaps with the previous and next images, a filter is used in order to consider only the landmarks that are within the 60 % of the center of the image.

B. Test and Results

Three different tests were carried out in order to prove the feasibility and robustness of the learning and recognition approach:

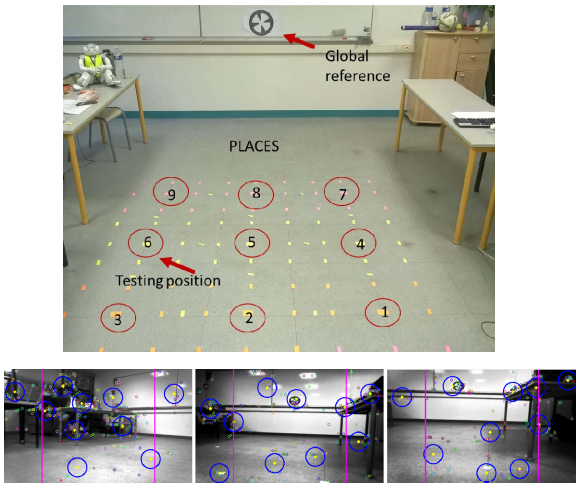

Figure 5. **Top part**: Test environment composed of nine different places. An artificial landmark is used as a global reference to compute the position of the landmarks. **Bottom part**: Three overlapping images extracted by the robot's camera. Only the landmarks in the center are considered.

TEST #1: To begin with, a single place in the environment was chosen to be learned and tested. The idea was to prove if after being learned it could easily be recognized not only when the robot was positioned at the same place but also when the robot was positioned at any place within the environment test.

RESULT #1: When the robot was at the exact same place where it had previously been, the corresponding **place cell** was activated with the higher possible value: 1, whereas as the robot went away from the place, the recognition value decreased proportionally to the distance (see Fig.6.1).

TEST #2: In this test, all the nine places depicted in fig. 5 were first learned. Then, the robot was placed at only one of them and the activity of all places was computed.

RESULT #2: Figure 6.2 shows the activity of all places when the robot was positioned on the 6th place. As expected, the activity of the corresponding place cell got the maximum value. Since the other places were previously learned, the corresponding place cells were activated, although with a minimum activation value.

TEST #3: This test consisted in verifying if the robot could learn different places as it navigates the environment in an incremental way.

RESULT #3: Whenever, a place was not recognized the robot would automatically learn it as a new one. Indeed, when the place's activity was below a given threshold, a vigilance term mechanism was activated indicating that the perceived place had not been recognized and needed to be learned. Figure 6.3 illustrates the activity probability of the place cells as the robot gets far from the learned "place 6 in orange". Then, since the value is little compared to the threshold value, it learns the place as a new one: "place 3 in blue".

In all tests, when the robot was placed at the exact same location of any of the learned places, the recognition rate was of 100 %. While the results of the first and second tests allowed us to validate the recognition of one or several places after having been learned, those of the third test allowed us to

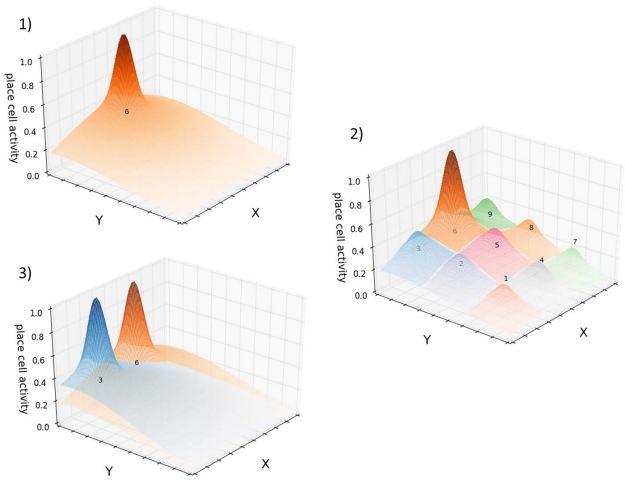

Figure 6. 1) Place "6" learned and recognized. 2) All nine places are learned. The high activation of place 6 indicates that the robot is positioned at that place and it recognized it as such. 3) Place "6" is learned and recognized, then as the robot navigates the environment, the activity value is inferior to the threshold value and the robot learns place "3" as a new one.

validate the incremental learning as the robot navigated the environment. This is an essential property for an autonomous navigation allowing the robot to adapt itself to the environment.

However, even though the recognition of the learned places was ensured at a 100% when the robot happened to be positioned at the exact learned places, different results showed that the recognition of the place dropped drastically when the robot's position was slightly modified with respect to the tested one. Since, the recognition of the landmarks was accurate; we suppose that the limitation comes from the determination of the threshold value and possibly the management of the lateral diffusion explained previously. This problem opens new perspectives for the improvement of the proposed algorithm.

VI. CONCLUSIONS

A biologically inspired approach for recognizing places within an environment has been presented in this paper as a robust solution for an autonomous robot navigation. The presented paper shows the robustness of the approach when navigating in complex environments where unforeseen situations are prone to happen such as the occlusion or absence of an expected sign serving as reference in the navigation task. Experimental results show the efficacy of the implemented algorithm. The robot is able to detect by itself what it considers to be relevant from the environment as natural landmarks and learns them as such. The ambiguity given by clusters with similar descriptors is dealt by the additional cluster position information which together are used to learn new places and recognize them during robot navigation. The learning task is performed in an incremental way allowing the robot to adapt itself to the environment.

Thanks to the generic composition of our previous neural architecture, the proposed model can be injected into it to further develop it with respect to robustness and completeness.

REFERENCES

[1] Gaussier, P., Revel, A., Banquet, J. P., Babeau, V. "From view cells and place cells to cognitive map learning: processing stages of the hippocampal system," Biological Cybernetics 86, pp. 15-28 (2002).

[2] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[3] Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Transactions on Neural Networks, 3, 698–713. Carpenter, G. A., Grossberg, S., &.

[4] D. M. Rojas Castro, A. Revel, and M. Menard. "Document image analysis by a mobile robot for autonomous indoor navigation." Document Analysis and Recognition (ICDAR), 2015 13th International Conference on IEEE, 2015.

[5] D. M. Rojas Castro, A. Revel, and M. Ménard. "A Robust Neural Robot Navigation Using a Combination of Deliberative and Reactive Control Architectures." European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2015. Page 445-450.

[6] Cummins, M. and Newman, P., "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," in [The International Journal of Robotics Research], 27, 647–665, SAGE Publications (June 2008).

[7] Ho, K. and Newman, P., "Detecting loop closure with scene sequences," International Journal of Computer Vision (IJCV) 74(3), 261–286 (2007).

[8] Eade, E. and Drummond, T., "Unified loop closing and recovery for real time monocular slam," in [British Machine Vision Conference (BMVC)], (2008). Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[9] Ramos, F., Upcroft, B., Kumar, S., & Durrant-Whyte, H. (2012). A Bayesian approach for place recognition. Robotics and Autonomous Systems, 60(4), 487-497.

[10] Wu, J., Christensen, H., & Rehg, J. M. (2009, October). Visual place categorization: Problem, dataset, and algorithm. In Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on (pp. 4763-4770). IEEE.

[11] Nguyen, V. A., Starzyk, J. A., & Goh, W. B. (2013). A spatio-temporal Long-term Memory approach for visual place recognition in mobile robotic navigation.Robotics and Autonomous Systems, 61(12), 1744-1758.

[12] Redish, A. and Touretzky, D. "Cognitive maps beyond the hippocampus". Hippocampus 7 (1): 15-35 (1997).

[13] Filliat, D., Meyer, J.-A. . "Global Localization and Topological Map Learning for Robot Navigation," in From Animals to Animats 7 Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior, edited by Hallam et al., The MIT Press, pp 131-140 (2002).

[14] C.R. Gallistel, The Organization of Learning, MIT Press, Cambridge, MA, 1993.

[15] S.P.D. Judd, T.S. Collett, Multiple stored views and landmark guidance in ants, Nature 392 (1998) 710–712.

[16] O'Keefe J, Dostrovsky J. 1971. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Res. 34:171–75.

[17] Gaussier, P., Joulain, C., Banquet, J., Lepretrê, S., & Revel, A. (2000). The visual homing problem: An example of robotics/biology cross fertilization. Robotics and Autonomous Systems, 30, 155–180.

[18] Giovannangeli, C. and Gaussier, P. (2008). Autonomous vision-based navigation: goal-orientated planning by transient states prediction, cognitive map building, and sensorymotor learning. Proceedings of the International Conference on Intelligent Robots and Systems, Nice, France, 2008.

[19] Ungerleider, Leslie G., and James V. Haxby. "'What'and 'where'in the human brain." Current opinion in neurobiology 4.2 (1994): 157-165.