Building Specific Contexts for On-line Learning of Dynamical Tasks through Non-verbal Interaction

Antoine de Rengervé, Souheïl Hanoune, Pierre Andry, Mathias Quoy, Philippe Gaussier

Building Specific Contexts for On-line Learning of Dynamical Tasks through Non-verbal Interaction

Antoine de Rengervé, Souheil Hanoune, Pierre Andry, Mathias Quoy and Philippe Gaussier
ETIS, CNRS ENSEA University of Cergy-Pontoise, F-95000 Cergy-Pontoise, France
{rengerve,souheil.hanoune,andry,quoy,gaussier}@ensea.fr

***Abstract*—Trajectories can be encoded as attraction basin resulting from recruited associations between visually based localization and orientations to follow (low level behaviors). Navigation to different places according to some other multimodal information needs a particular learning. We propose a minimal model explaining such a behavior adaptation from non-verbal interaction with a teacher. Specific contexts can be recruited to prevent the behaviors to activate in cases the interaction showed they were inadequate. Still, the model is compatible with the recruitment of new low level behaviors. The tests done in simulation show the capabilities of the architecture, the limitations regarding the generalization and the learning speed. We also discuss the possible evolutions towards more bio-inspired models.**

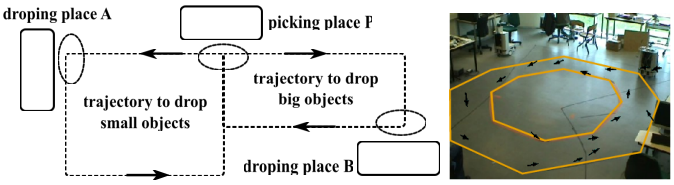

Fig. 1. *Left* Example of contextual task in navigation. A solution is to encode contexts biasing the selection of actions in the particular situations (locations (circles) + objects status) when there is a choice to make. In place P, the decision is between picking objects, moving left and moving right. In place A and B, the possible actions are dropping held object and moving to picking place. *Right* Example of typical trajectory learned through interaction as an attraction basin emerging from place-cell/orientation couples (black arrows).

I. Introduction

Action selection is the choosing of the most appropriate action out of a set of possible candidates. The word "action" can represent notions ranging from high level abstracts ("pour water in a glass") to low level motor commands ("move arm joint with chosen speed"). We are interested in a task related to the second case (Fig. 1). A robot must navigate to different places depending on the transported object. The robot takes an object at the picking place P and goes to the correct dropping places (A or B) to release the objects depending on their sizes. The robot has to select the adequate actions (moving directions i.e. low level actions) according to the current sensory inputs. After the mid 80's, solutions to action selection problem based solely on executing the steps of a given plan to achieve a goal have been progressively abandoned for more reactive behaviors [1]. Using behavior modules that always directly monitor the environment allows the system to react faster to changes. The organization of these modules can be in parallel [2] or in hierarchy [3]. It has also been argued that partial activation of reactive modules depending on a carefully designed hierarchy with attentional switching can also provide a robust solution with an easier coordination of modules [4]. In [2], the behavior modules are encoded as a condition (a combination of sensory inputs determining when the behavior can activate), an action (what the behavior consists in) and a result (expected sensory inputs after the action is performed). At first, the condition is general and the activation of the behavior is easy. Learning increases the selectivity of the condition to match the fact that the desired result cannot be obtained by the particular action in any situation. The learning is based on the correlation between the occurring of particular sensory inputs and the occurring of the result after the action is executed. Sensory inputs are progressively integrated in the condition of a behavior to better determine when it should be active, improving the coordination between the different behaviors. Reinforcement learning is another candidate solution for the action selection problem [5]. Relying on neurobiological results, computational models of the Basal Ganglia were developed to explain the action selection by humans. The properties of dopamine neurons indicate that they could implement some kind of reinforcement learning [6]. The Basal Ganglia, organized in parallel loops (potentially coding actions) with mutual inhibition, can perform the competition necessary for the action selection [7], [8]. However, both the reinforcement learning and the correlation learning are quite slow to learn. In order to have a reactive interaction, the behavior should be adapted in a faster way.

Following a simple trajectory can be encoded as an attraction basin emerging from multiple place-cell/action associations [9]. A place is encoded by merging "what" is seen with "where" it is seen[1]. A recruited place-cell (PC) responds accordingly to the distance to the encoded place with a maximal response when at the learned spot (see [9] for details). The actions are direction of movement[2]. The robot selects its moving direction depending on its place in the environment (see an example of learned trajectory in Fig. 1). The obtained navigation controller is robust due to the generalization capabilities of the place-cell recognition. New place-cell/action couples are learned from on-line non-verbal interaction [10]. When the robot moves too far from the desired trajectory, the robot is shown how to get back to the desired trajectory by forcing its orientation[3]. This corrected orientation is associated to a new learned place-cell completing the encoded attraction basin. In order to learn the task of Fig. 1, the information about

[1]The codes use visual, proprioceptive (camera orientation) and magnetic compass information to build place-cells.

[2]Orientations of movement are given with respect to an absolute reference (North). The orientation of the robot is read from a magnetic compass.

[3]In the experiments of this article, a joystick was used considering that it simulates the action of a leash.

the object (e.g. size, but it could be visual, tactile,. . .) should be included in the condition of the actions i.e. transforming place-cells into multimodal categories. Two approaches are possible to build the adequate multimodal categories. Local solutions to action selection can be learned and progressivelly adapted to enable good generalization. For instance, when the behavior of the robot is corrected, multimodal categories can be recruited and associated with the correct actions. The generalization depends on how the different modalities contribute to the context activation. Without any particular a priori, the recruited categories would include all the modalities equally. Thus, the generalization properties should be improved by learning how much each modality should contribute, with a possible pruning of the irrelevant links. The alternate solution is to start from a general category (i.e. taking only a few modalities into account) and progressively increase the selectivity of the category (like with conditions in Maes's model [2]). Selectivity can be adapted on the basis of the feedback provided by the interaction. In the framework of the place-cell/action based controller, we propose that the place-cells be non-specific categories progressively refined by inhibiting them in situations when they predict undesired behaviors according to the interaction from the teacher. Some multimodal contexts are thus recruited to store which place-cells should be inhibited and when. A task may not be solved by only refining the existing PC/action couples. The recruitment of new PC/action couples may occur to enrich the basis of behaviors of the system. We will also show that the whole process can run in parallel with the aforementioned trajectory learning.

In Section II, we present the neural network based architecture recruiting contexts for the inhibition of irrelevant PC/action couples. The evaluation of the actions and the context recruitment are detailed with the conditions to recruit new place-cells. The model is implemented in the neural network simulator *Promethe* [11], and tested in a robotic simulation based on Webots (Cyberbotics) (Sec. III). The first experiment studies the behavior of the learning process when the contexts are not needed to learn a trajectory. Then, the model is validated in the task described in Fig. 1. The model manages to learn the task. Even though this learning scheme shows limited generalization properties, we discuss in Sec. IV whether they could be a basis for longer learning that will focus on summarizing the contexts into chunks. Such a fast adaptation of the action selection may be useful to complete and even train a slower learning network extracting the statistics of the task.

II. Model for specific context building based on interactive learning

The combination of several PC/action couples is sufficient to shape an attraction basin so that the robot follows a desired trajectory (Fig. 1 and 2, [9]). The action evaluation block in the architecture learns the contexts in which some place-cells should be inhibited. The place-cells are thus biased by the output I of the action evaluation process before the competition between the biased place-cells. The winning biased place-cell PC^I determines the selected action i.e. the orientation to be followed. Figure 3 details the action evaluation process performed in the block shown in Fig. 1. Indeed, an action is encoded as the dynamics which maintains a particular orientation of movement. During an interaction, the teacher

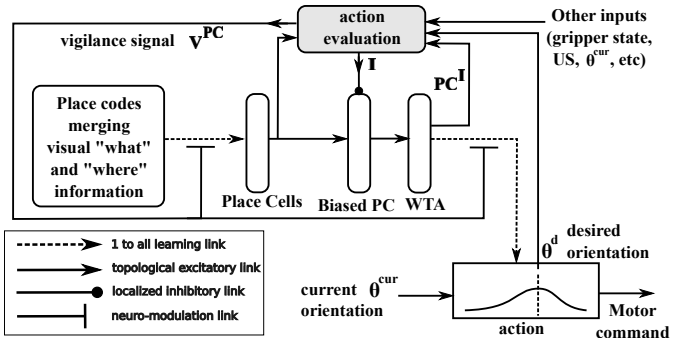

Fig. 2. Architecture for place-cell based navigation with the action evaluation part of the architecture (see Fig. 3). Sensory and predicted signals are provided to the action evaluation system. It outputs a vigilance signal and an inhibitory signal preventing some place-cells to exhibit their associated orientation. The vigilance signal can trigger the learning of a new place-cell.

corrects the behavior of the robot by opposing to the predicted dynamics i.e. by moving the current orientation away from the current desired orientation. The interaction may even be more like making the robot turn left or right than explicitly giving the exact direction to follow (no explicit supervision). Selected PC/action couples predicting rotations opposite to the executed rotation during the interaction are to be inhibited. Contexts are created to encode which place-cells should be inhibited and when. After this learning, if the behavior happens to be corrected again, the task has probably changed. Therefore, when a new interaction phase starts in an already known context, the previous associations with place-cells to inhibit will be removed to learn the new situation. However, a particular case may be taught by the teacher by alternating between correcting behaviors and evaluating the result. Though, it will be sensed by the robot as separated interaction phases. The aforementioned process is completed by a short term memory of the wrong PC/action couples to ensure the consistency of the teaching. Even if, the contextual associations are reset prematurely, the short term memory keeps the results of recent detections. Finally, repeated corrections in the same context mean the condition refining of existing behaviors fails. Hence, a new PC/action couple is recruited adding a new behavior to solve the task, possibly through condition refining.

A. Wrong action detection

The neural layer D^W outputs the result of the detection. The neurons in D^W match the existing place-cells. An activity of 1 in D^W means that the corresponding place-cell should be inhibited because the associated action is evaluated as wrong. A competition between the different biased place-cells PC^I determines the action to be performed (Fig. 2). Currently, only the action predicted by the winning biased place-cell $i_M = \underset{i}{argmax}(PC_i^I)$ can be evaluated. Therefore no wrong action detection can succeed with other PC/action couples ($D^W_{i \neq i_M} = 0$). In the case of the i^{th}_M couple, the action is evaluated as wrong if, during the interaction, the robot orientation θ^{cur} moves away from the robot proposed orientation θ^d. The teacher's control being primary, he can prevent the robot from following the desired orientation θ^d. The sensorimotor error E_r is based on the difference between these two orientations $E_r = min(\mid \theta^d - \theta^{cur} \mid, (\theta_{max} - \theta_{min}) - \mid \theta^d - \theta^{cur} \mid)$ with

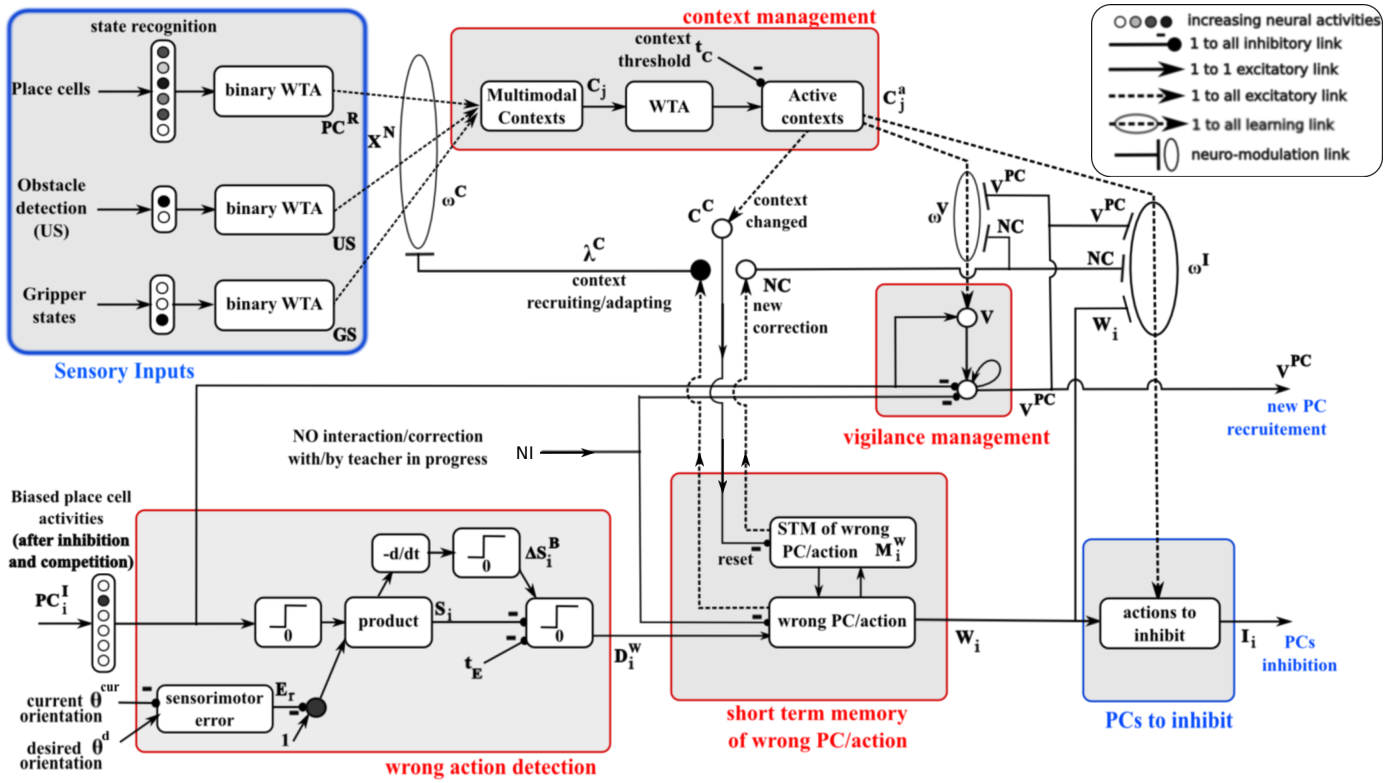

Fig. 3. Context learning to inhibit PCs associated to wrong actions and learn new place-cells through on-line interactive teaching. Given a correction, the variation of the similarity is used to estimate if the selected action is wrong. A multimodal context is then recruited if none already encodes the situation. Each time a new interaction phase starts, the actions to inhibit are reset and the contextual vigilance increases. The short term memory of wrong actions ensures the consistency of a split interaction. Eventually, if the activity of the winning place-cell after inhibition PC^I is under the vigilance V, a new place-cell/action couple will be recruited. This architecture details the content of the action evaluation block shown in Fig. 2.

$\theta_{max} = 2\pi - \epsilon$ and $\theta_{min} = 0$. This particular computation is needed because the orientation space loops on itself (2π rad is the same as 0 rad). The similarity measure S for the i^{th}_M PC/action couple uses this error to recognize how much the current orientation is similar to the desired orientation (Eq 1).

$$S_{i_M} = 1 - \frac{2E_r}{(\theta_{max} - \theta_{min})} \tag{1}$$

with $\mathcal{H}_0$ the Heaviside function that verifies $\mathcal{H}_0(0) = 0$. The coefficient multiplying E_r normalizes the dynamics of the similarity measure (values between 0 and 1). Detecting that the $i_M{}^{th}$ PC/action couple is wrong ($D^W_{i_M} = 1$) depends on the dynamics of the similarity ($\Delta S^B_{i_M}$) as well as on its value S_{i_M} (Eq 3). An action is estimated as wrong if the evolution of the similarity indicates that the followed orientation moves away from the desired one (Eq 2) and if the similarity is low enough. When a negative variation of the similarity for the $i_M{}^{th}$ action is detected, $\Delta S^B_{i_M}$ is equal to 1. In that case, the similarity measure is compared with a similarity threshold equal to $1 - t_E$, with t_E equivalent to an error threshold.

$$\Delta S^B_{i_M}(t) = \mathcal{H}_0(S_{i_M}(t - \Delta t) - S_{i_M}(t)) \tag{2}$$

$$D^W_{i_M} = \mathcal{H}_0(\Delta S^B_{i_M} - t_E - S_{i_M}) \text{ and } D^W_{i \neq i_M} = 0 \tag{3}$$

B. Context management

Contexts are recruited when the behavior of the robot is corrected. The input X of the multimodal contexts C is the concatenation of the different discrete binary codes for each sensory modality. The raw place-cell activities PC^R give localization information. The obstacle detection US based on ultrason sensor is categorized into 2 neurons (frontal obstacle present or not). The gripper state GS is encoded by three neurons each one corresponding to an opening width. The input X is normalized and then connected to the context layer C (Eq 4). The context learning (Eq 5) is based on Adaptive Resonance Theory [12]. The maximal activity in the context layer C is compared with a vigilance threshold λ^C. If the maximum is lower, then a new context (with index r) is recruited so that the weights ω^C_{rl} reproduce the input pattern. As the input is normalized, the activity of a context is maximal when the same encoded pattern is presented again.

$$X^N_k = \frac{X_k}{|| X ||} \text{ with } X = [PC^R; US; GS] \tag{4}$$

$$\begin{cases} C_j = \sum_k \omega^C_{jk} \cdot X^N_k \\ \Delta\omega^C_{rk} = (X^N_k - \omega^C_{rk}) \text{ if } \lambda^C > \max_j(C_j) \end{cases} \tag{5}$$

with λ^C the vigilance threshold equal to 0.99 if there is an active neuron in the wrong action layer W and 0 otherwise. The active context layer C^a only contains one active neuron corresponding to the maximally recognized context $j^M = \underset{j}{argmax}(C_j)$ in C. The activity of the winning context must be over the context threshold t_C (Eq 6).

$$C^a_{j^M} = \mathcal{H}_0(C_{j^M} - t_C) \text{ and } C^a_{j \neq j^M} = 0 \tag{6}$$

C. Short term memory of wrong PC/action couples

During an interactive teaching phase, the PC corresponding to detected wrong actions are memorized in a short term memory. The wrong PC/action layer W depends on the new detected wrong action D_i^W as well as on the content of the memory of the recent incorrect actions M^W (Eq 7). If there is no ongoing interaction, the content of the wrong action layer is inhibited. Otherwise, any active neuron in this layer will trigger the learning of a multimodal context C and the association between this context and the neurons corresponding to the same PC in the inhibition layer I.

$$W_i = \mathcal{H}_0(M_i^W + D_i^W - 2 \cdot NI) \tag{7}$$

The short term memory M^W is fed by the wrong action layer W. The temporal forgetting γ ensures that the memory is reset within a few seconds. It can also be reset when the index of the winning context in C^a changes ($C^C = 1$).

$$M_i^W(t) = [M_i^W(t - \Delta t) - \gamma]^+ + W_i - 2 \cdot C^C \tag{8}$$

with $[x]^+ = x$ if $0 < x < 1$, 0 if $x < 0$ and 1 otherwise. A new correction phase starts whenever one neuron in the wrong actions memory becomes active. During the first iteration a new correction phase is detected ($NC = 1$). The phase ends when all neural activities decay to 0 with the forgetting or when a reset signal is received because the context changed ($C^C = 1$).

D. Learning PCs to inhibit

The inhibition layer I contains the PCs to inhibit. The associations are learned with a Hebbian like rule (Eq 9) and stored on the synaptic weights ω^I connecting the active context layer C^a with the layer I. When a new correction starts ($NC = 1$), the previously stored inhibitions are reset. When the teacher is correcting the robot, the wrong PC/actions stored in the short term memory M^W transit through the wrong PC/action layer W. The PCs to inhibit are associated with the active context in I. The context/PC associations are also reset when a new place-cell is recruited ($\alpha^{PC} = 1$).

$$\begin{cases} I_i = [W_i + \sum_j \omega_{ij}^I \cdot C_j^a]^+ \\ \Delta w_{ij}^I = W_i \cdot (C_j^a \cdot I_i - \omega_{kj}^I) - (NC + V^{PC}) \cdot C_j^a \cdot \omega_{ij}^I \end{cases} \tag{9}$$

where $V^{PC} = 1$ when a new place-cell is recruited. The PCs represented in the layer I are inhibited so that the selected orientation is predicted by one of the correct place-cells (see Figure 2).

E. Learning new PC/action couples

The aforementioned process may fail because PC/action couples with sufficiently recognized PCs may not be already encoded. The vigilance threshold V^{PC} controlling the learning of new PC/action couples is thus managed (Eq 10). Each time a new correction occurs in a context ($NC = 1$), the vigilance V associated with this context increases.

$$\begin{cases} V = [\sum_j \omega_j^V C_j^a]^+ \\ \Delta\omega_j^V = NC(PC_{i_M}^I - V) \cdot C_j^a - V^{PC} \cdot V \cdot C_j^a \end{cases} \tag{10}$$

Fig. 4. Simulated environment and robot. The robot has to learn to navigate between the different blocks according to the object size signal.

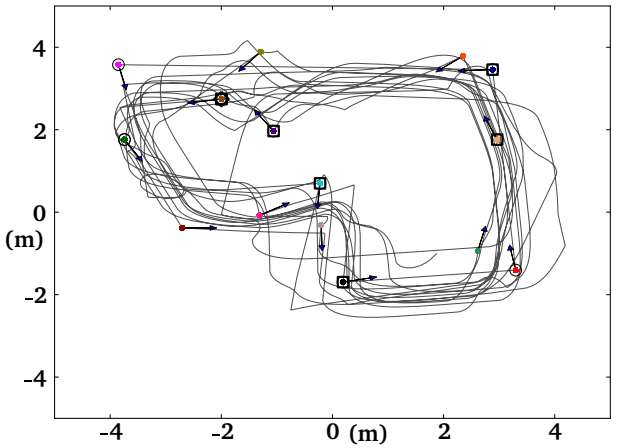

Fig. 5. Trajectory (gray line) during learning and reproduction of the task (several times consecutively). The colored dots are the learned place-cells. The chunks associated with the place-cells are represented around the corresponding dots (circle: the context exists but does not inhibit an action, square: the context does not inhibit the winning PC, star: the context is used). The arrow starting from the dots indicates the orientations that are associated to the place-cells. At the end of the learning, only one chunk still modifies the behavior. The others only inhibit PC that could not win.

with $i_M = \underset{i}{argmax}(PC_i^I)$. The activity of the winning action PC^I is compared to this vigilance threshold $V^{PC} = \mathcal{H}_0(V - PC_{i_M}^I)$ to trigger the recruitment of a new place-cell. When a new place-cell is recruited, the context/PC association for the context in which occurred the recruitment is reset ($V^{PC} = 1$ in Eq 10). The context/action association in the inhibition layer I is also reset for this context ($V^{PC} = 1$ in Eq 9). The system can switch between the PC inhibition and learning new PC/action couples.

III. Emergent attraction basins and contexts

In this section, the model is tested in two different experiments performed with a simulated robot, running with the Promethe simulator under the Webots (Cyberbotics) 3D environment. The robot is composed of a mobile platform with a camera mounted on a pan servomotor (Fig. 4). The walls of the room are covered with pictures to provide textured images for visual processing i.e. the recognition of the place-cells. In the first experiment, the robot has to learn a simple trajectory that does not require multimodal contexts. As the new model extends the model from [10], the initial test aims at studying whether the recruited contexts are useful and whether the system will rely on these contexts to tackle the task. In Figure 5, the learned place-cell/orientation associations are

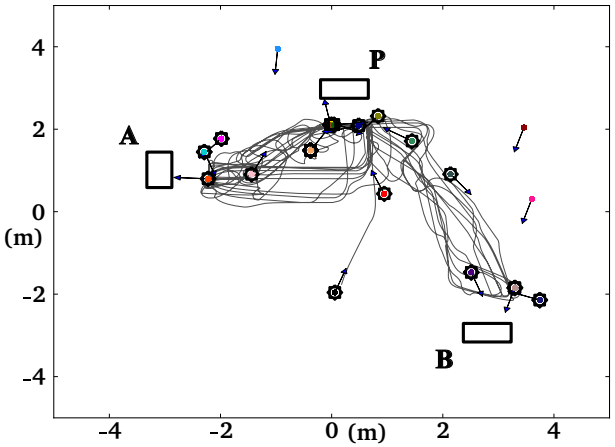

Fig. 6. Trajectory (gray line) during reproduction of the task (several times consecutively). The colored dots are the learned place-cells. The chunks associated with the place-cells are represented around the corresponding dots. The arrows starting from the dots indicate the orientations that are associated to the place-cells. Due to the effect of the chunks, the orientation may not be followed as the PC selecting this orientation is inhibited.

displayed with the followed trajectories. During the learning phase, the first time a correction is performed in a given place-cell. A context is recruited to encode the sensory configuration. Each time the orientation of the robot is corrected, the active context is associated with the inhibition of the incorrect PC/orientation couples. When a new place-cell is recruited in a context, previous associations between this context and the PCs to inhibit are unlearned. At the end of the learning phase, i.e. when the robot is finally capable of performing rounds without correction, the contexts do not influence the behavior anymore because they do not inhibit the already winning PC/orientation couple (see Fig. 5). The contexts appear as only temporarily used. Only one context still influences the behavior. It corresponds to the situation where the robot detects a wall nearby, and thus should aim at a different direction than the one associated with the winning place-cell.

In the second experiment, we focus on testing the learning of a task that requires selecting the trajectory depending on sensory information not directly related to navigation. The information is the width opening of the gripper i.e. the size of the held object. The environment contains three obstacles at interesting locations (see Fig. 4). The white one is where the robot has to take objects, the two others (red and blue) are where the robot can drop them. The colors of the obstacles are not directly used by the system. The task of the robot is to take objects at the picking place and then, depending on the simulated size of the object, move to one location or the other to place it there. The objects are simulated through the gripper state coding the object size. A correct orientation maintained during a few seconds is a prerequisite to validate the reaching to one of the places. In this experiment, there is no obstacle avoidance behavior so that the robot can face an obstacle without avoiding it. Yet, the forward speed of the robot is decreased when the robot gets closer to an obstacle so that it will not bump into it. The robot is corrected by simply changing its orientation (using a joystick) whenever it moves in the wrong direction. The trajectory followed by the robot during the reproduction of the task is presented in Figure 6. The learning phase is quite long as each new place-cell introduces a new potential context preventing previous generalization from other context and also a new possible PC/orientation that may have to be inhibited when in other contexts. The robot can reproduce the task without any correction after about fifteen rounds of the pick-and-place task using a small object and eleven rounds using a big object. As the robot only needs correction when it makes mistakes, less and less corrections are necessary as the learning process goes on.

At the end of the learning process, the robot performs the task without any correction. The learned contexts enable the robot to exhibit the correct behavior. Depending on the gripper information, the robot follows different trajectories. Even if the learned contexts are very specific and encode particular sensory configurations, the effective attraction basins can be interpreted as depending on the gripper state. The resulting attraction basins are displayed in Figure 7. Because the robot did not start from any location in the environment, the attraction basin is only correct in a limited area. Depending on its starting position, the robot can end up stuck in front of a wall (as there is currently no obstacle avoidance). The robot may also select the wrong dropping place. For instance in Fig. 7b), the dropping place on the left is where the robot should go with a small object. However, the dropping place on the bottom-right of the figure is also an attractor where the robot can go depending on its starting position. Without the corresponding learning, the robot has generalized the dropping place on the bottom-right as valid for any size of object. This is due to the fact that the learned PC/orientation couples in the vicinity of this place converge to the dropping place. The interest of the contexts appears when no object is held, by reducing the attraction basin so the robot can move to the picking place. Hence, the robot exhibits the expected contextual behavior that can emerge from the recruited contexts and the learned inhibition of PCs.

IV. Discussion

In this paper we presented a model of action selection based on contexts guiding sensory-motor associations. This approach allows us to extend the place-cell/action model [9] to solve the action selection problem. In [10], the authors showed the influence of the teacher on the learned trajectory by comparing different methods of learning. The authors showed that a compromise between proscriptive and prescriptive learning was more robust and that the choices of the human interacting teacher corresponded to such compromise. We expect the same kind of influence to be at work in the contextual navigation task. The issue will be explicitly addressed in future work, in order to validate that the convergence of the context learning is not (too) dependent on the expertise of the teacher. Besides, in our experiment, we tested the model in a simulated environment, without variations and always in the same situation.

The goal of this paper was to study a minimal model that could explain the fast adaptation of multimodal behaviors during interaction with a teacher. The results showed that the proposed principles are efficient but need to be improved for a real autonomous development. First, the contexts are computed with a threshold that prevent generalization on different place-cells. Second, we also used binary values to build the sensory inputs of the contexts. This allowed us to ease the study but

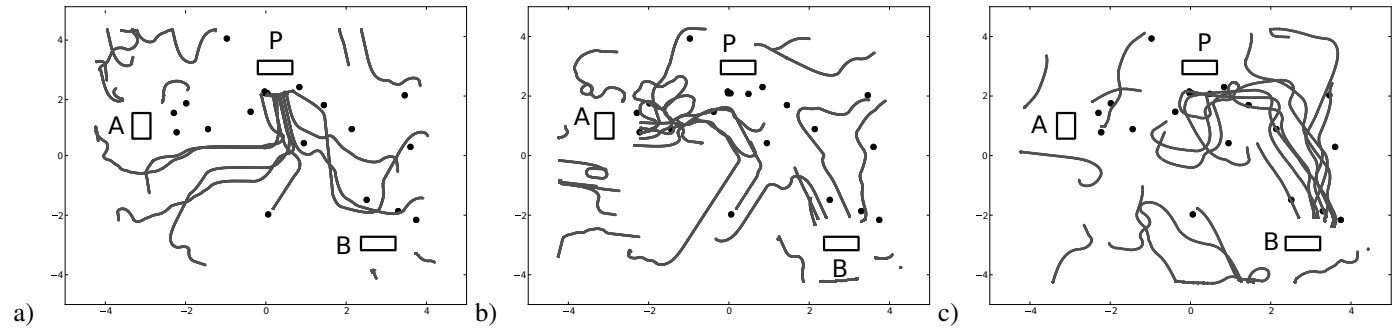

Fig. 7. Change in behavior according to the held object. Trajectories (gray) from different starting points when the gripper is holding (a) no object, (b) small object, (c) big object. Contextual attraction basins toward the pickup place P, the dropping place A and the dropping place B can emerge even though the contexts encode specific sensory configurations.

this constraint should be removed for better generalization. The generalization properties of the existing low level behaviors also determines the quality of overall behavior. There should be a cooperation between the proposed fast learning of contexts and a slower learning solution improving the low level behaviors at the basis of the system.

Many neurobiological studies were dedicated to the action selection process and thus can give hints to improve the capabilities of a robotic system. Different cerebral regions are involved in the action selection process. In particular, the implication of the cortico-baso-thalamo-cortical loop was clearly exposed [13]. The GPR model [7] is based on a dynamical system approach and the internal connectivity of the BG is unknown. The CBTC model [8] tends to improve the exposure of the internal connectivity and introduces a fusion with reinforcement learning mechanisms to improve the GPR model. Those models tend to be very complex. In this paper, we focused on a minimal solution to obtain an action selection behavior, by using a selective inhibition of PC/action couples depending on the recruitment of multimodal contexts. However, this fast on-line learning is to be completed by a slower learning that could encode chunk like categories directly selecting the action to be done. The theory of chunking was first introduced in the 1950s by DeGroot [14] and Miller [15]. The main idea is that a chunk collects pieces of information in order to obtain a higher level of information coding. In a previous work [16], we suggested that a modified version of Schmajuk's and DiCarlo's learning of conditioning [17] could model the cortico-basal loop with associative conditioning in the cerebellum and resulting in the learning of chunks. Our goal is now to combine the fast on-line learning of contexts presented in this paper with the aforementioned slower learning of chunks to improve the action selection capabilities of the robot.

ACKNOWLEDGMENT

This work was supported by the AUTO-EVAL project, the INTERACT french project ANR-09-CORD-014, and the NEUROBOT project ANR-BLAN-SIMI2-LS-100617-13-01.

REFERENCES

[1] R. Brooks, "A robust layered control system for a mobile robot," *IEEE J. Robot. Autom.*, vol. 2, no. 1, pp. 14–23, 1986.

[2] P. Maes and R. Brooks, "Learning to Coordinate Behaviors," in *AAAI Proceedings*, 1990, pp. 796–802.

[3] T. Tyrrell, "Computational mechanisms for action selection," Ph.D. dissertation, University of Edinburgh., 1993.

[4] J. J. Bryson, "Hierarchy and Sequence vs. Full Parallelism in Action Selection," in *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*, J. A. Meyer, A. Berthoz, D. Floreano, H. Roitblat, and S. W. Wilson, Eds. Cambridge, MA: MIT Press, 2000, pp. 147–156.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 1998.

[6] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.

[7] K. Gurney, T. J. Prescott, and P. Redgrave, "A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour." *Biol Cybern*, vol. 84, no. 6, pp. 411–423, Jun. 2001.

[8] B. Girard, D. Filliat, J.-A. Meyer, A. Berthoz, and A. Guillot, "Integration of navigation and action selection functionalities in a computational model of cortico-basal-ganglia-thalamo-cortical loops," *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, vol. 13, no. 2, pp. 115–130, Jun. 2005.

[9] C. Giovannangeli, P. Gaussier, and G. Désilles, "Robust mapless outdoor vision-based navigation," in *IEEE/RSJ International Conference on Intelligent Robots and systems*. Beijing, China: IEEE, 2006.

[10] C. Giovannangeli and P. Gaussier, "Interactive teaching for vision-based mobile robots: A sensory-motor approach," *IEEE Trans. Syst., Man, Cybern. A*, vol. 40, no. 1, pp. 13–28, 2010.

[11] M. Lagarde, P. Andry, and P. Gaussier, "Distributed Real Time Neural Networks In Interactive Complex Systems," in *proceedings of the IEEE International Conference on Soft Computing as Transdisciplinary Science and Technology (CSTST 08)*, 2008, pp. 95–10.

[12] G. A. Carpenter and S. Grossberg, "Adaptive resonance theory (ART)," in *The handbook of brain theory and neural networks*. Cambridge, MA, USA: MIT Press, 2002, pp. 79–82.

[13] T. J. Prescott, J. J. Bryson, and A. K. Seth, "Introduction. modelling natural action selection," *Philosophical Transactions of the Royal Society B - Biological Sciences*, vol. 362, no. 1485, pp. 1521–1529, 2007.

[14] A. D. De Groot, *Thought and Choice in Chess*. Mouton De Gruyter, 2nd edition (June 1978), 1978.

[15] G. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review*, vol. 63, no. 2, pp. 81–97, 1956.

[16] S. Hanoune, M. Quoy, and P. Gaussier, "An architecture for online chunk learning and planning in complex navigation and manipulation tasks," in *IEEE International Conference on Development and Learning (ICDL) and on Epigenetic Robotics (Epirob)*, 2012, pp. 1–6.

[17] N. A. Schmajuk and J. J. DiCarlo, "Stimulus configuration, classical conditioning, and hippocampal function," *Psychological Review*, vol. 99, no. 2, p. 268–305, apr 1992, PMID: 1594726.