Proceedings of the 2008 IEEE
International Conference on Robotics and Biomimetics
Bangkok, Thailand, February 21 - 26, 2009

Curiosity-Driven Acquisition of Sensorimotor Concepts Using Memory-Based Active Learning*

Sergio Roa, Geert-Jan M. Kruijff, and Henrik Jacobsson
Language Technology Lab
German Research Center for Artificial Intelligence / DFKI GmbH
Saarbrücken, Germany
{*sergio.roa,gj, henrik.jacobsson*}*@dfki.de*

Abstract—**Operating in real-world environments, a robot will need to continuously learn from its experience to update and extend its knowledge. The paper focuses on the specific problem of how a robot can efficiently select information that is "interesting", driving the robot's "curiosity." The paper investigates the hypothesis that curiosity can be emulated through a combination of active learning, and reinforcement learning using intrinsic and extrinsic rewards. Intrinsic rewards quantify learning progress, providing a measure for "interestingness" of observations, and extrinsic rewards direct learning using the robot's interactions with the environment and other agents. The paper describes the approach, and experimental results obtained in simulated environments. The results indicate that both intrinsic and extrinsic rewards improve learning progress, measured in the number of training cycles to achieve a goal. The approach presented here extends previous approaches to curiosity-driven learning, by including both intrinsic and extrinsic rewards, and by considering more complex sensorimotor input.**

Index Terms—**Intrinsically motivated reinforcement learning, interactive robot learning, developmental robotics, epigenetic robotics.**

I. Introduction

We would like our robots to operate in real-world environments. They should assist us at home, in the office, malls and supermarkets, or even outdoors. The challenge we face there is that these are perceptually very rich environments. It is not possible to endow the robot with all there is to know about such environments – we can in no way guarantee "omniscience out of the box." To address this challenge, we need to make the robot capable of learning what it does not know yet. Using its experience, it should continuously update and extend its knowledge.

In this paper we address a specific problem in the wider context of continuous robot learning. As the robot is able to obtain rich perceptual input, how can it efficiently obtain and select information that is relevant or interesting, given what it is trying to learn? The hypothesis we explore in this paper is that a robot can focus on interesting learning material by adopting an active form of exploration. We propose a combination of active learning, with reinforcement learning based on intrinsic and extrinsic rewards. The intrinsic rewards focus on learning progress, whereas the extrinsic rewards direct learning based on the robot's interactions with the environment and other agents. The combination of the rewards provide the basis for the robot's curiosity and motivation to explore some aspects of the environment further.

The approach we present in this paper combines insights from active learning mechanisms for speeding up learning [1], and intrinsic motivation systems for learning [2]. Intrinsic motivation systems drive the learning process by measuring how and whether learning makes progress. In addition to intrinsic motivations, our approach also includes various extrinsic sources of motivation. Extrinsic sources include the robot's interactions with the environment, and other agents.

The intrinsic motivation system we adopt is based on the idea of successive stages of development. We use a fusion of the perception and action state spaces to define a sensorimotor model. At each time step, a tuple consisting of the current sensorimotor space and the perceptual state estimation is stored, i.e. a prediction of the consequences of the actions can be evaluated. We then use the error in prediction to calculate a measure of learning progress. The intrinsic rewards used to guide the exploration process are inversely proportional to the decrease in error rate of the experts used for prediction. Effectively this means that the opportunities to learn more are triggered by this mechanism in an active manner. As basis for our prediction models we use the memory-based KD-tree algorithm for k-nearest neighbour search[3]. However, more sophisticated learning machines should be used for larger training times, in order to lower the space computational complexity of the problem.

Learning successively enters into more complicated stages of development by sequentally splitting the sample space of sensorimotor and prediction features at specific time steps. In cycles of 250 time steps, the sample space is split based on its variance, defining cutting values to divide the sample space. Thus, after some number of splits, prediction machines become distributed, and start to concentrate on the specific state space regions for which they yield optimal predictions.

We designed the extrinsic motivation system to include rewards based on interaction with the environment, and with a human tutor. The later type of rewards is inspired by the

*The research reported of in this paper is supported by EU FP6 IP "CoSy" #4250

experiments of [4]. Types of extrinsic rewards are collision penalties, evaluation of progress towards a physical goal object, and rewards on succesfully executing particular actions (i.e. when the robot succesfully grips an object). Moreover, the human tutor is able to interact with the robot by sending a reward signal whenever he/she finds this appropriate. We maintain a memory of the more recent rewards that are relevant on the current sensorimotor context.

Given our hypothesis, the expectation is that intrinsic and extrinsic rewards help a robot to achieve an expected goal in a shorter amount of time. To test this, we created a simulated world (using Player/Stage), and placed a mobile robot in a room in a larger environment. In the environment, the robot encounters obstacles that it should be able to avoid or surpass. We defined several experiments, differing in task complexity and in what types of rewards were available to the robot. In each experiment, the goal task is for the robot to grip an object. Depending on the scenario, this goal object may be in a room different from the one in which the robot starts. As robot we use a Pioneer P2-DX equipped with sonar sensors, fiducial and blob finders, bumpers and a gripper touch sensor. These sensors are sources of information that can be detected by the robot as salient events, for instance when directing the attention to objects. The space of motor actions the robot can perform is continuous and allows three degrees of freedom, i.e., it is able to move backwards and forwards, rotate left and right, and close or open the gripper.

In this paper we show results for three different kinds of experiments. In the first experiment, only intrinsic motivation is employed. The second one applies also extrinsic motivation but not interactive rewards, and the third one includes also interactive rewards. These experiments provide indications that the use of rewards based on learning progress is indeed beneficial for reaching the goal object. Adding extrinsic rewards can help accelerate reaching the goal (again, on the average) by a 60% – or, combined, intrinsic and extrinsic rewards can help cut cycles by 80% over a standard active learning baseline. The experiments also demonstrate the effectiveness of the intrinsic motivation system when applied to a sensorimotor space which is more complex than those explored in related work. The robot manages to explore efficiently the environment, and explore regions of the learning space that might be interesting while avoiding situations of low learning progress.

This paper is organized as follows. The next section presents a brief overview of current research in intrinsically motivated systems and interactive learning. In section III, we explain the learning algorithm in detail. In section IV we present and discuss the experimental results. We close the paper with conclusions and discussions of follow-up research.

II. Related work

The opportunities for exploration and curiosity have been found to be important mechanisms for animal, humans and robots to learn (see discussion in [5], [2]). There exists a kind of intrinsic motivation system which is a source of internal rewards, in contrast to extrinsic rewards that can be obtained from the environment or other external agents. Thus, such a system rewards exploration without the need of immediate external rewards. The discovery of a new skill is then a reward in itself. For children it is in fact more important to autonomously explore the world to gain motor and perceptual abilities in its first stages of development, although an adult teacher can help by scaffolding the children's environment [6]. This learning process is also active, in the sense that the opportunities to learn more interesting tasks are progressively chosen by the agent. Different motivation drives have been considered, such as novelty, surprise, incongruity and complexity.

Two scenarios were set up in [2] to evaluate this learning mechanism. In the first experiment, a simulated robot equipped with two wheels predicts a distance to a toy based on the consequences of taking some actions and the distances it senses. The action space is three dimensional and consist of the speed of motor on the left, on the right and the frequency of a sound emission. Depending on three different frequency ranges, in this simulated environment the toy moves either randomly, or it stops, or jumps into the robot. The actions to choose are selected according to the maximization of the expected reward (inverse of error rate decay) in the next time step, allowing also a random exploration of actions with a given probability. In this work, typical reinforcement learning algorithms such as *Q-Learning* were not considered in order to avoid the complex issues arising from the consequences of delayed rewards. However, these techniques are applied in related works [7] and in general this approach is commonly known as intrinsically motivated reinforcement learning. It is important to notice that rewarding learnability punishes predictability and complex unpredictability. These aspects have been also considered in [8].

A second experiment involves a Sony AIBO robot in interacting with toys that can be bitten, bashed or visually detected. Its sensors can perceive the detection of an object, the ocurrence of biting an object, and the toys oscillations. No a priori knowledge of the consequences of actions was included in the robot programming, apart from proper control primitives needed for perception. In this more difficult experiment, the results also show that the robot usually starts performing random actions, followed by simple tasks and finally more complex ones. When it finds an interesting source of learning it spends some time performing the corresponding action, till there is no more motivation for doing it. Thus, the robot recognizes affordances, that is, that certain actions are sufficiently interesting at some point when the robot identifies the correlations between these actions and its corresponding perceptions.

The results of the experiments demonstrated that the agent starts performing actions almost uniformly randomly and then it focuses on more and more complicated stages, where the actions consequences depend on more variables. Thus, the robot avoids situations in which nothing can be learned and directs its attention autonomously to more complex situations. It was also shown that this algorithm (Intelligent Adaptive Curiosity - IAC) is more efficient than simple random exploration or IAC without exploration.

In [6], [4], an active virtual agent called Sophie learns the order of steps needed in cooking. These skills are obtained not only by trial and error tests but also by interaction with a human teacher, which is able to give feedback as well, based on sending a reward signal. Experiments conducted with human tutors (not machine learning experts) demonstrate that they are able to identify when the agent is making learning progress, when some feedback mechanism (*transparency*) is included in the agent behaviour.

III. Approach

As pointed out in section I, the robot interacts in its environment by using a curiosity-driven behaviour mechanism. We developed an Intrinsic Motivation System, which is based on the work described in [2]. In order to implement a curiosity-driven motivation mechanism, we want the robot to concentrate on situations that are new or interesting for it. Thus, following the work in [2], the robot is able to predict the tuple $\{\mathbf{SM}(t-1), \mathbf{S}(t)\}$, where $\mathbf{SM}(t)$ is the concatenation of the sensor and motor vectors $\mathbf{S}(t)$ and $\mathbf{M}(t)$ at time t. We use learning machines to predict the consequences of taking some action in a given sensory state at the previous time step. In the first time steps, a learning machine corresponding to a first region R_1 is created. All sensorimotor perceptions found are considered to be part of this region. In our experiments, a region is split into 2 regions after 250 time steps, as described in [2]. The sensorimotor context is partitioned by using a measure of the variance of the instances in the region, and a cutting value and cutting index in the sensorimotor space is used as splitting criterion. In this way, this information-based procedure is used to partition efficiently the state space. Thus, the learning machine M_n, corresponding to a region R_n, specializes in some sensorimotor context.

The error rate $e_n(t)$ is tracked for successive time steps in order to measure an average error rate. The decrease in error rate is obtained and this quantity is used to calculate the learning progress, that is, an error rate reduction corresponds to an increase in learning progress. The intrinsic reward $r_n^l(t)$ is then the calculated learning progress quantity.

As previously sketched, we also make use of extrinsic rewards in order to guide the robot to reach the goal. These are a penalty for collisions r_n^c, a reward for the gripping event r_n^g and a reward r_n^f when an approximation to a distance goal is measured by using the fiducial finder. Moreover, the interactive reward mechanism r_n^{int} can be employed by the human tutor. The overall reward mechanism is then:

$$r_n(t) = \sum_i \alpha_i r_n^i(t), \tag{1}$$

where α_i is the weight of the i-th reward applied in the region R_n. In our experiments, we gave similar weights for all the rewarding techniques, except for the collision penalties that get lower values. When the robot collides with the goal object, we notice that this situation is also interesting from the point of view of the goal we want the robot to achieve.

The learning procedure is performed as follows. A first region is created and the previous sensorimotor state $\mathbf{SM}(t-1)$ is registered, together with the current sensory state $\mathbf{S}(t)$. Then, a learning machine is used to learn this training instance. The sensorimotor vector is normalized with values ranging from 0 to 1. Afterwards, the vectors $\mathcal{H}_n = \{\mathbf{C_n}, \mathbf{G_n}, \mathbf{I_n}\}$ of recent rewards are updated according to the sensing information, where $\mathbf{C_n}, \mathbf{G_n}$ and $\mathbf{I}_n$ correspond to collision, gripping and interactive rewards respectively. In this case, a history of 15 events is stored. For each reward vector $\mathbf{K}_n \in \mathcal{H}_n$, the corresponding current reward r_n^i is calculated as:

$$r_n^i(t) = \sum_{t \geq t_n} \varphi^{t-t_n} K_n^{t_n}, \tag{2}$$

where φ is a discount factor, typically 0.99. The fiducial finding based reward r_n^f is calculated as a sum of the differences between the successive x and y distances to the goal object.

Then, a new action should be selected that maximizes the expected rewards. For this purpose, a sample of 100000 possible actions is generated and the expected learning progress $L_n(t) \approx r_n^l(t-1)$ and expected extrinsic rewards $E\{r_n^i(t)\} \approx r_n^i(t-1)$ for the current Region R_n are calculated. So, the maximum value for all generated actions $\mathcal{A}$ for some regions $\mathcal{R}$ is calculated:

$$max\_r(t) = \arg\max_{\mathcal{A},\mathcal{R}} \sum_i \alpha_i r_n^i(t-1) \tag{3}$$

for some region $n \in \mathcal{R}$. The corresponding action is executed by the robot. When selecting translational actions, we assure that the robot does not perform dangerous actions like approaching hastily the walls. Moreover, we used a near(ϵ)-greedy action selection rule with $\epsilon = 0.3$ to allow random actions and permit additional exploration sources. Thus, a random action is selected with a probability of 0.3.

IV. Evaluation

A. Evaluation setup: Scenarios & methods

In this work, the PlayerStage simulator was employed to perform experiments. A Pioneer P2-DX robot was included in the world, plus two objects that are used as obstacles and a last object to be gripped by the robot. This object owns a fiducial that is detected by the robot when the object stands in the robot vision range. The robot acts in a room environment surrounded by walls and a hole to pass to a next room (see Figure 1).

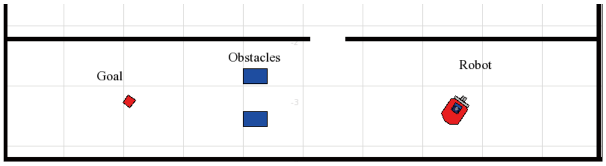

Fig. 1. The scenario in the initial state. The robot (right-hand side) is expected to surpass the obstacles in the middle and reach and grip the goal object (left-hand side).

At each time step in the learning loop, the robot senses its environment using the following sensors:

- 16 sonars distributed around the robot, whose values range is $[0.0, 5.0]$.
- 3 bumpers at right, left and front sides of the robot, with values 0 and 1, the latter corresponding to a collision detection.
- 2 gripper sensors, corresponding to a touch sensor and a gripper status.
- 1 fiducial finder sensor detecting x and y relative coordinates from the robot to the detected object.
- 1 blob finder sensor detecting area and object position in the x coordinate in the vision range of the robot.

Thus, a sensor context $\mathbf{S}(t)$ measured in time step t is a 25-dimensional continuous vector, since the robot is able to detect at most 3 objects at a given time using the blob finder.

The motor context $\mathbf{M}(t)$ in time step t is a 3-dimensional continuos vector of:

- translational velocity values ranging in the interval $[-0.5, 0.5]$.
- rotational velocity values ranging from $-0.1\frac{180}{\pi\omega}$ to $0.1\frac{180}{\pi\omega}$, where ω is the wheel diameter of the robot, where $\omega = 24$ for a Pioneer P2-DX. These quantities were found to allow more stability in the robot behaviour.
- 3 gripper motor values corresponding to the actions open, close and no action.

Three different settings are considered in this investigation. In the first place, only intrinsic motivation mechanisms are used. Secondly, extrinsic motivations are also included and finally also interactive rewards. Moreover, different scenarios where the state space of actions and sensors is restricted were set up, in order to analyse more carefully the whole system by studying its different components.

In each of these cases the learning progress is taken into account, because the system is mainly based on this curiosity-driven mechanism. When the robot performs actions corresponding to a specific region R_n, we observe increases and decreases of the learning progress.

In order to analyse the different sources of reward independendently, we performed several experiments on a restricted state space. The first scenario involves the robot moving in circles around its own axis and a gripable object. With this setting (*gripping scenario*), we want to check that the robot enters in a stage of learning progress, when it finds the gripping action interesting, i.e., when succesfully gripping the object. The second scenario involves a robot in front of the object at a certain distance. In this setting (*reach goal scenario*), we want to test the effectiveness of the extrinsic rewarding techniques in order to reach the goal.

B. Results

As explained in the previous section, we made preliminary analyses of the data by restricting the state space of the robot. In the *gripping scenario*, the robot only performs rotational and gripping actions. In Figure 2 we can observe a correlation between the learning progress and the increase in the frequency of actions that are interesting for the robot, e.g., closing the gripper to grip the object.

The *reach goal scenario* involves a target object and the robot performs only translational actions in front of it at a certain distance. Figure 3 shows that the use of fiducial finding based rewards and interactive rewards accelerates the task of approaching the object. 5 different runs of the experiment were performed for 3 different cases. In average, $\sim$ 100 steps are needed to reach the goal for the scenario without extrinsic and interactive rewards; $\sim$ 40 for the scenario with extrinsic rewards and $\sim$ 20 with extrinsic and interactive rewards.

The last scenario, whose initial state is the one we show in Figure 1 is evaluated with the whole motivation system. In Figure 4, one possible path taken by the robot is shown. We found that when using additionally extrinsic rewards, the average number of steps needed for the robot to attain the goal ($\sim$ 3000) does not decrease. This might be due to the fact that the fiducial rewards are not always accesible, because there are different ways to surpass the obstacles. The other rewards are sometimes not relevant enough for reaching the goal. This was also the case for interactive rewards and the reason might be that given that each reward is related to a specific region and this region might correspond to a rich sensorimotor state (when different actions are allowed), this reward might be causing unexpected effects for actions we want to reward and not to punish. This mechanism is, however, useful when the robot gets stuck at certain regions that have been well learned after some time, for instance the regions bordering walls. Figure 5 shows the contours of some perceptions of the robot until it reaches the goal object.

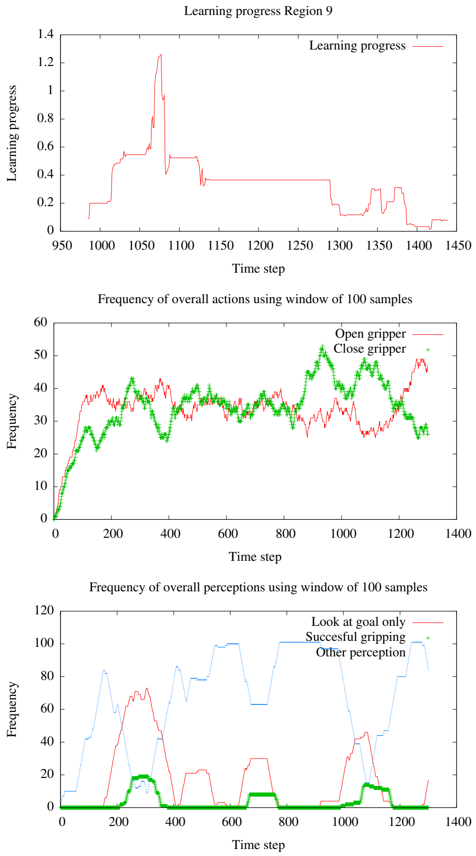

Fig. 2. The robot manages to grip the object and close the gripper at a higher frequency than the average at the time step ∼ 1050. The learning progress in the corresponding context region 9 increases. By "other perception" we mean bordering walls, looking at obstacles, among others.

C. *Result analysis*

The results show the effectiveness of the active learning mechanism to explore the environment and recognize interesting sources of learning. However, the results presented here are preliminary and we have to perform more analyses in order to understand the nature of the learning progress. Since the perceptions of the robot change quickly because of the nature of the motion, it is more difficult to establish when the creation of successive stages of development is relevant for achieving some goal. For instance, some regions are created but the robot perform actions very rarely in such state spaces. Additionally, sometimes the robot escapes quickly from a recent created region.

In spite of this, observations of the learning curves show that the robot in fact learns from the environment. One example is when the robot manages to reach the obstacles by going forward and backwards, detecting a salient event. We have observed peaks in these actions when the robot finds these sources of learning.

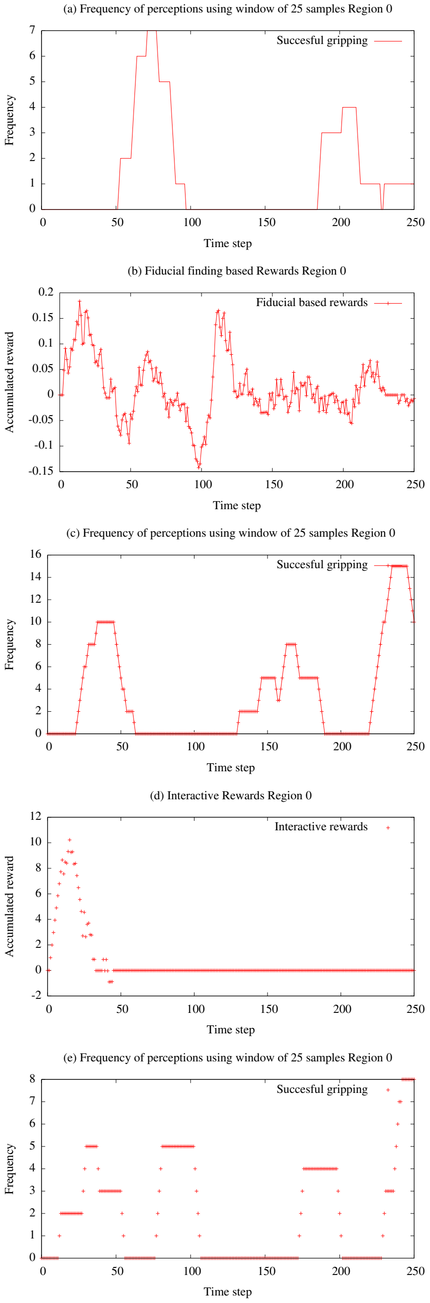

Fig. 3. Figure (a) shows the gripping event by only using learning progress reward. Figures (b) and (c) show that the fiducial based rewards permit the robot to reach the goal in ∼ 25 time steps. (d) and (e) show that the interactive rewards accelerate more the goal reach. The memory of 15 rewards also allows the robot to reach the goal more frequently.

There are also some issues related to the effectiveness of the reward mechanism. Since the rewards are related to a specific region, it is possible that when some regions generalize over many different types of actions the rewarding mechanism might be counterproductive. In the next section, we discuss different approaches to these problems.

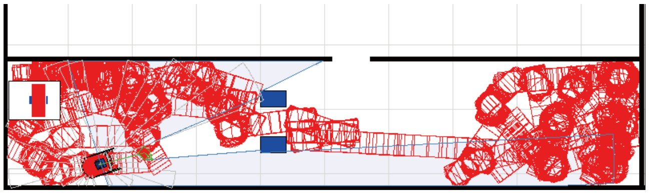

Fig. 4. A path taken by the robot using intrinsic motivation. In this figure, it is observed that the robot first focus on trying to learn the region around the walls, and then identifies a salient event such that the obstacles and reach them. After this, it also finds the goal twice and experiments with the walls in the left in between. The box at the left-hand side of the image shows the blob finder sensor data.

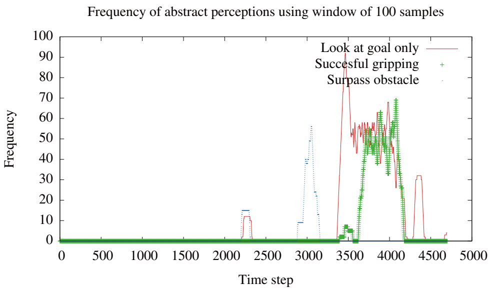

Fig. 5. One run of the experiment, when we observe some abstract perceptions of the robot in the scenario with obstacles and gripable objects

Because of some limitations in the speed simulation allowed by the software, we were not able to perform as much experiments as needed, which is also important to obtain additional data to be analysed more carefully.

V. Conclusions and Discussion

In this paper, we describe a curiosity-driven based mechanism for exploration of a mobile robotic environment. Regions of interest arising from the partition of the state space are successively created, allowing the robot to select proper actions given a specific sensor context. Interesting results were found and it is recognized that the robot is able to explore and learn from the environment using an intrinsic mechanism. Moreover, other external sources of rewards were also investigated, as well as interactive rewards, which are found to be sometimes useful when the robot gets stuck at some not interesting regions or to accelerate the approximation to a target objective.

The robot is able to attain goals, i.e., to reach some obstacles, surpass them and then reach and grip a goal object. However, much work remains to be done in order to understand the robot behaviour and improve the rewarding techniques.

Moreover, it is also an open issue how the agent can retain temporal information and use hierarchical mechanisms to abstract simple tasks into more complex ones, when it is put in an autonomous setting. This has been studied for reinforcement learning configurations specially using the concept of *options* for the generalization of tasks [9], [6], [10], [11].

It might be useful to explore alternatives like exploration in specific state spaces using some kind of invariance [8]. Moreover, one can consider automatic construction of Markov models after an efficient exploration using intrinsic motivated approaches [7], [12], [13], which use more sophisticated reinforcement learning algorithms like Q-Learning but have not been tested in real robotic environments. Moreover, prediction of motivational drives or rewards has also been investigated and other sources of motivation like predictability or familiarity may also be taken into account [8], [6], including more complex motivational systems.

References

[1] D. Cohn, L. Atlas, and R. Ladner, “Improving generalization with active learning,” *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[2] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, “Intrinsic motivation systems for autonomous mental development,” *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 1, 2007.

[3] J. H. Friedman, J. L. Bentley, and R. A. Finkel, “An algorithm for finding best matches in logarithmic expected time,” *ACM Transactions on Mathematics Software*, vol. 3, no. 3, pp. 209–226, September 1977.

[4] A. L. Thomaz, G. Hoffman, and C. Breazeal, “Reinforcement learning with human teachers: Understanding how people want to teach robots,” in *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2006.

[5] P.-Y. Oudeyer and F. Kaplan, “Intelligent adaptive curiosity: a source of self-development,” in *Proceedings of the 4th International Workshop on Epigenetic Robotics*, L. Berthouze, H. Kozima, C. G. Prince, G. Sandini, G. Stojanov, G. Metta, and C. Balkenius, Eds., vol. 117. Lund University Cognitive Studies, 2004, pp. 127–130. [Online]. Available: citeseer.ist.psu.edu/oudeyer04intelligent.html

[6] A. L. Thomaz, “Socially guided machine learning,” Ph.D. dissertation, Massachusetts Institute of Technology, May 2006.

[7] A. Stout, G. Konidaris, and A. Barto, “Intrinsically motivated reinforcement learning: A promising framework for developmental robot learning,” in *Proceedings of the AAAI Spring Symposium on Developmental Robotics*, Stanford University, Stanford, CA, March 21-23 2005.

[8] F. Kaplan and P.-Y. Oudeyer, “Motivational principles for visual know-how development,” in *Proceedings of the 3rd Epigenetic Robotics workshop : Modeling cognitive development in robotic systems*, ser. Lund University Cognitive Studies, C. Prince, L. Berthouze, H. Kozima, D. Bullock, G. Stojanov, and C. Balkenius, Eds., vol. 101, 2003, pp. 72–80.

[9] R. S. Sutton, D. Precup, and S. P. Singh, “Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning,” *Artificial Intelligence*, vol. 112, no. 1-2, pp. 181–211, 1999. [Online]. Available: citeseer.ist.psu.edu/sutton99between.html

[10] J. Provost, B. Kuipers, and R. Miikkulainen, “Self-organizing distinctive state abstraction using options,” in *Proceedings of the 7th International Conference on Epigenetic Robotics*, L. Berthouze, C. G. Prince, M. Littman, H. Kozima, and C. Balkenius, Eds., vol. 135. Lund University Cognitive Studies, 2007.

[11] ——, “Developing navigation behavior through self-organizing distinctive state abstraction,” *Connection Science*, vol. 18, no. 2, 2006.

[12] S. Singh, A. G. Barto, and N. Chentanez, “Intrinsically motivated reinforcement learning,” in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 1281–1288.

[13] Özgür Şimşek and A. G. Barto, “An intrinsic reward mechanism for efficient exploration,” in *ICML ’06: Proceedings of the 23rd international conference on Machine learning*. New York, NY, USA: ACM, 2006, pp. 833–840.