Incremental Skill Acquisition for Self-Motivated Learning Animats

Andrea Bonarini, Alessandro Lazaric, and Marcello Restelli

Department of Electronics and Informatics
Politecnico di Milano
piazza Leonardo da Vinci 32, I-20133 Milan, Italy,
{bonarini,lazaric,restelli}@elet.polimi.it,
WWW home page: http://www.elet.polimi.it

Abstract. A central role in the development process of children is played by self-exploratory activities. Through a playful interaction with the surrounding environment, they test their own capabilities, explore novel situations, and understand how their actions affect the world. During this kind of exploration, interesting situations may be discovered. By learning to reach these situations, a child incrementally develops more and more complex skills. Inspired by studies from psychology, neuroscience, and machine learning, we designed SMILe (Self-Motivated Incremental Learning), a learning framework that allows artificial agents to autonomously identify and learn a set of abilities useful to face several different tasks, through an iterated three phase process: by means of a random exploration of the environment (*babbling phase*), the agent identifies interesting situations and generates an intrinsic motivation (*motivating phase*) aimed at learning how to get into these situations (*skill acquisition phase*). This process incrementally increases the skills of the agent, so that new interesting configurations can be experienced. We present results on two gridworld environments to show how SMILe makes it possible to learn skills that enable the agent to perform well and robustly in many different tasks.

1 Introduction

In this paper, we describe SMILe (Self-Motivated Incremental Learning), a learning framework leading an agent to incrementally learn general abilities through a direct interaction with the environment guided by self-generated interest. This approach integrates ideas coming from *cognitive sciences* and *intrinsically motivated reinforcement learning* and defines a self-development process that enables animats to autonomously operate in complex environments.

In recent years, studies on the inner mechanisms of human development, pursued in many different areas (such as neuroscience, psychology, developmental sciences, robotics, machine learning) converged to a new field, commonly referred to as *developmental robotics*[7, 20]. Traditionally, a designer must specifically program the set of skills needed for an animat to accomplish a given task. Often, these skills are tuned to perform a predefined task on a specific environment, and the learned abilities can hardly be reused if the task or the environment changes. On the other hand, developmental robotics tries to reproduce the basic mechanisms at the basis of human and animal development processes so as to propose frameworks in which the agent does not directly address any specific problem, but develops a set of basic skills up to very general abilities that can be used to solve many different tasks.

Because of the complexity of its goal, developmental robotics has many different facets [7]. In this paper, we focus on a subset of them and we will consider the developmental process as an *incremental process* where an agent *organizes* its initial skills through *spontaneous exploratory phases* and *self-motivated learning activities*. Self-motivated learning proved to be one of the most challenging aspects of development processes, as shown in [19, 2, 8]. One of the most promising approaches is *intrinsically motivated reinforcement learning* [2], that enables an agent to autonomously develop a hierarchy of skills through a process guided by an intrinsic motivation, without any commitment to achieve a specific task.

The SMILe framework extends the intrinsically motivated reinforcement learning model to a more general development process, in which the notion of interest is not hardwired, but autonomously extracted from characteristics of the environment. The learning process of each skill has been decomposed into three phases (*babbling*, *motivating* and *skill acquisition*), that are endlessly iterated to develop a hierarchy of abilities that can be exploited by animats to better control the environment.

The rest of the paper is organized as follows. In the next section we give a general description of SMILe and we introduce a novel framework for self-motivated learning. Section 3 gives an overview of the implementation of the framework using Reinforcement Learning (RL) techniques, and we provide a general definition of the interest function. Section 4 provides some experimental results on two gridworlds that simulate simple robotic environments, showing how the acquired skills help the agent to reduce the learning time in many different tasks. Finally, in the last section we discuss the results and propose some possible future directions.

2 The Learning Process

As stated in [20], one of the most promising approaches to achieve the ambitious goal of autonomy in artificial systems, is the definition of a suitable lifelong development process. This consists of an open-ended learning process in which an agent pursues self-motivated goals and develops highly reusable skills. Developmental robotics has its main source of inspiration in studies from neuroscience and psychology [7], that show how similar mechanisms could be traced in the developmental process of children.

Many approaches in developmental robotics refer to the studies by Piaget [13], and to his research on children's early stages of development. Piaget showed that childish development can be considered as an incremental process of acquisition

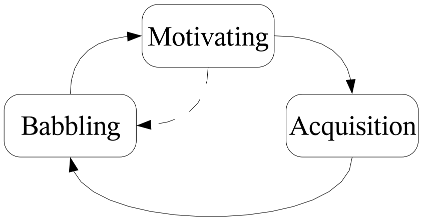

Fig. 1. The self-motivated developmental learning process of SMILe

of new abilities in which children modulate the complexity of their activities in association with the increasing complexity of their cognitive and morphological structures. Another important contribution to the comprehension of the mechanisms attending human development comes from the research carried out by Berlyne [3] about the notion of curiosity and its influence on behavior and the rising of intrinsic motivation. Berlyne asserts that, in absence of a particular aim, human behavior is partly determined by an innate will of exploring what is perceived as interesting. Psychologists define curiosity as a form of motivation that promotes exploratory behavior to learn more about a source of uncertainty.

In summary, life-long learning seems to be characterized by a progressive, self-motivated *development* that leads to the *incremental* acquisition of more and more complex skills. SMILe implements this concept into a simplified learning process suitable for artificial systems, whose aim is to incrementally learn new skills that could be potentially useful to face different tasks. Each skill is learned by a self-motivated process that iterates on three main phases (see Fig. 1):

- *Babbling*: the agent playfully interacts with the environment to get aware of the relationships between its actions and the environment dynamics.
- *Motivating*: the agent evaluates which is the most interesting situation it has experienced during the exploration performed during the babbling phase.
- *Skill acquisition*: the agent learns the skill to reach the interesting situation.

2.1 Babbling Phase

One of the crucial activities in the development process of puppies and babies is *self-exploration* [7]. Through self-explorative acts, they become aware of their own capabilities with respect to the surrounding environment, understand the consequences of the actions they have autonomously selected, and learn to control and exploit the dynamics of their bodies. In analogy to vocal babbling, this experiential process has been called *body babbling* [10].

Moving from these observations, we introduced in SMILe, at the beginning of each iteration, a *babbling* phase. The acquisition of a new skill starts with a self-explorative phase in which the agent, for a certain time, randomly executes its admissible actions. The choice of taking actions according to a uniform probability distribution over the action space, even if it is not fully compliant to body babbling theory, is consistent with the fact that this exploration is completely goal free, without any external motivation leading the agent behavior. The goal of the babbling phase is to collect information about the environment dynamics that can be used in the next phase to determine whether there is any *interesting* skill that is worth learning.

2.2 Motivating Phase

There is a huge body of evidence about the central role played by *intrinsic motivation* in development process as the main driver of organisms behavior when no extrinsic motivation is available. In this way, they may increase their competence to control the environment, by acquiring a broad set of skills that can be reused for different goals. Studies from psychology, like those of Piaget [13] and Berlyne [3], and from neuroscience [5], suggest that intrinsic motivation may be generated by several factors: surprise, incongruity, and novelty. All these factors act together to determine an intrinsic *interest* associated to different situations.

Many studies [15, 12] relate interest to the current knowledge of the observer and its capability to predict the outcome of its interaction with the environment, and propose particular quantitative definitions of novelty and surprise.

In SMILe, this second phase computes the *interest function*, which associates an interest value to each state visited during the babbling phase. Despite previous approaches, we propose a general methodology to define and compute interest values for each state obtained by the propagation (through the estimated transition model of the environment) of a given local measure of interest (a formal definition will be given in Section 3). Once the interest function has been computed, SMILe determines the next goal by searching for the state associated to the maximum interest. Learning to reach this state is the goal of the third phase.

It may happen that the agent has no strong motivation in learning to reach a state rather than another. In this case, it makes no sense to spend time and efforts in learning something that is not so interesting, but it is better to start a new babbling phase in order to collect more experience that could allow to discover new interesting situations (represented by the dashed line in Fig. 1).

2.3 Skill Acquisition Phase

During a development process an organism starts with very simple skills and acquires more and more complex abilities. Each time a new skill is learned, it may be used to simplify the learning of the following ones, thus progressively increasing the complexity of the tasks that could be successfully faced.

Recently, the idea of hierarchically decomposing complex problems into simpler sub-problems has been successfully exploited also in RL with the introduction of formalisms for managing temporally extended actions [1]. Several of these approaches work with fixed hand-coded decompositions, even if some proposals have been advanced to dynamically decompose a given goal into simpler subgoals [9, 11]. Barto et al. [2] have proposed an intrinsically motivated approach to generate the hierarchy of skills.

In SMILe, during the skill acquisition phase the agent learns, through an intrinsic reward function, a skill that leads to the most interesting state identified by the motivating phase. While learning the skill, the agent, in addition to its basic actions, may benefit also from other previously acquired skills.

After the acquisition of a new skill, the development process of SMILe starts a new iteration activating a new babbling phase, in order to experience how the new skill modifies the agent interaction with the environment. This leads to the computation of a new interest function that defines a new learning goal, thus obtaining an incremental learning process that continuously increases the agent capabilities of controlling its environment.

3 SMILe

In this section we propose an implementation of the learning framework described in Section 2. As already proved in many studies [2, 21, 19], Reinforcement Learning (RL) is one of the most suitable frameworks to deal with learning problems in developmental robotics. Furthermore, the incremental development of simple skills into complex activities can be efficiently described using Hierarchical Reinforcement Learning (HRL) [1], as suggested in [2].

3.1 Formal Representation of Skills: the Option Framework

HRL problems are generally formalized using Semi-Markov Decision Process (SMDP) models. In particular, in the *option framework* [1] an SMDP is defined by tuple $\langle \mathcal{S}, \mathcal{O}, \mathcal{P}, \mathcal{R} \rangle$, where $\mathcal{S}$ is the set of states (i.e. perceptions), $\mathcal{O}$ is the set of options (i.e. skills), $\mathcal{P}(s, o, s')$ is the transition model, that is the probability to get to state s' taking option o is state s, and $\mathcal{R}(s)$ is the reward in state s. The main difference between traditional RL approaches and intrinsically motivated learning, concerns the source of reinforcement. While in the usual interaction model the agent receives a reinforcement signal provided by an external critic, we consider the reward as the result of an intrinsic motivation of the agent that pursues self-generated goals according to the model proposed in [2].

Formally, a skill is represented as an option o, i.e. a tuple $\langle \pi_o, \mathcal{I}, \beta \rangle$, where $\pi_o : \mathcal{S} \times \mathcal{O} \rightarrow [0, 1]$ is the control policy that describes the probability to execute an option when the agent is in a specific state, $\mathcal{I} \subset S$ is the set of states where the option is defined and $\beta(s)$ is the probability for an option to terminate at state s. When the development process starts, the agent has an initial set of basic options $\mathcal{O}^0$, at the k-th iteration, the set of options is incrementally modified adding the option learned in the skill acquisition phase: $\mathcal{O}^k = \mathcal{O}^{k-1} \cup \{o^k\}$.

3.2 Incremental Learning of Reusable Skills

In the following, we give a brief description of the implementation of the development phases of SMILe, summarized in Algorithm 1 (for more details see [4]).

In the *babbling phase*, at each time instant the agent simply executes one skill at random, choosing among the set of admissible skills $\mathcal{O}^k$. The aim of this phase is to build, at each iteration k, an estimate (even partial) $\widehat{P}_{\pi_R^k}(s, s')$ of the state transition probabilities when the random policy π_R^k is used for a sufficient number of steps. Since the state transition probabilities do not depend only on characteristics of the environment, but also on the abilities of the agent, when a new skill is learned, the capabilities of the agent to control the environment dynamics change and the state transition probabilities must be recomputed.

Through the playful exploration performed in the babbling phase, the agent experiences several different situations. In the *motivating phase*, SMILe computes the interest associated to each state on the basis of the information contained in the estimated state transition probabilities $\widehat{P}_{\pi_R^k}(s, s')$.

Algorithm 1 The SMILe Algorithm

```
repeat
 Babbling Phase
  for all Babbling episodes do
    for all Steps do
      Given state $s$, choose action $o$ at random over $\mathcal{O}^k$
      Take action $o$, observe state $s'$
      Update state transition probability estimation $\widehat{P}_{\pi_R^k}(s, s')$
    end for
  end for
 Motivating Phase
  Given model $\widehat{P}_{\pi_R^k}(s, s')$, compute local interest $\rho(s)$
  Compute interest function $I(s)$
  if no interesting state can be identified then
    step back to the Babbling Phase
  else
    Extract subgoal $s^* = arg \max_s I(s)$
    Create reward function $R(s)$
  end if
 Skill Acquisition Phase
  for all Skill Acquisition episodes do
    for all Steps do
      Given state $s$, choose action $o$ according to $\epsilon$-greedy
      Take action $o$, observe state $s'$ and reward $r$
      Update state-action value function
    end for
  end for
until forever
```

Although there are several characteristics of a model that could be used to compute the local interest of a state, such as transition entropy and controllability (details can be found in [14]), here we will focus on the following definition:

$$\rho(s) = (1 - p_{in}(s)) - p_{in}(s)\,(1 - p_{out}(s))\,, \tag{1}$$

where $p_{in}(s) = \frac{1}{|\mathcal{S}|}\sum_{s' \in \mathcal{S}} P_{\pi_R}(s', s)$ and $p_{out}(s) = \sum_{s' \neq s} P_{\pi_R}(s, s')$. The first term of Equation 1 is the probability of not moving into state s in one step following the policy π_R, given that the agent starts from a random state. To this term, we subtract a second term that represents the probability to reach s in one step starting from a random location and then to remain in s for another step. The intuition behind Equation 1 is that states that, under a random policy, are difficult to be reached or that, once reached, can be easily left, are relevant as subgoals for many complex tasks whose solution needs the agent to pass through states that cannot be easily reached without a specific skill.

This measure defines the concept of interest of a state only on the basis of information about its input and output transition probabilities, without taking into account the characteristics of the surrounding states; for this reason we call it *local interest functions*. Using the estimated state transition probabilities and a local interest function, we define the global interest function with the following Bellman-like equation:

$$I^k(s) = \rho^k(s) + \gamma \sum_{s' \in \mathcal{S}} \widehat{P}_{\pi_R^k}(s, s') I^k(s'). \tag{2}$$

In this way, the interest of a state depends, not only on the characteristics of its local transitions, but also on the interests of those states that may be reached from it. The discount factor $\gamma \in [0, 1)$ determines how much distant states should influence the interest of the current state. To compute $I(s)$ we can use an iterative policy evaluation algorithm that uses Equation 2 as an update rule [17]. The formulation of the interest function $I(s)$ given in Equation 2 is such that it can represent a large set of the aspects of the concept of interest depending on the specific definition of local interest $\rho(s)$ that is used.

Once $I^k(s)$ has been computed, the agent self-determines its next goal by choosing the most interesting state $\overline{s}^k = arg \max_s I^k(s)$, and produces an intrinsically motivated reward function that simply returns a positive reward when the agent achieves state $\overline{s}^k$ and null otherwise. It is possible to show that, using the definition of local interest previously introduced, the acquisition of new skills decreases the interest in goal states (*boredom effect*), thus preventing the agent from choosing them again.

As stated in Section 2.2, after some iterations, the interest function tends to flatten until no state with relevant interest can be identified in the motivating phase. In this case, the agent has no advantage from learning to reach new useless goals and the babbling phase is started again in order to either refine the transition model estimation or adapt to changes in the environment dynamics [4].

After having identified the goal state $\overline{s}^k$ and generated the intrinsically motivated reward function $\mathcal{R}^k(s)$, the agent starts the *skill acquisition* phase in

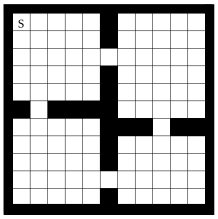

Fig. 2. Four Rooms world.

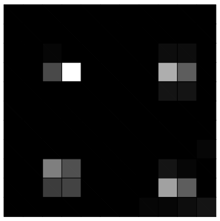

Fig. 3. Density plot of subgoal distribution in Four Room world.

Algorithm	*Mean*	*Max*
Q-Learning	$1.293 \cdot 10^4$	$3.569 \cdot 10^4$
Random	$1.756 \cdot 10^4$	$6.777 \cdot 10^4$
SMILe	$0.957 \cdot 10^4$	$2.044 \cdot 10^4$

Table 1. Average and maximum number of learning steps in Four-Rooms with random goal

which it learns the policy of a new option o^k whose goal is $\overline{s}^k$. The policy of the new option is learned according to the option learning algorithm described in [1]. At each time step, the action value function $Q(s, o)$, that is the estimation of the amount of reward the agent can obtain by taking option o in state s, is updated according to the following update rule:

$$Q(s,o) \leftarrow (1-\alpha)Q(s,o) + \alpha \left[\tilde{r} + \gamma^i \max_{o' \in O^{k-1}} Q(s',o') \right] \tag{3}$$

where α is a learning step size, i is the number of steps taken by option o to meet its termination condition and $\tilde{r}$ is the reward accumulated from s to s' in i steps according to the reward function $\mathcal{R}^k(s)$.

Once the skill acquisition is finished, the new option o^k is created and added to the set of options $\mathcal{O}^{k-1}$. This new option is characterized by a deterministic policy that can be directly derived from the action value function $Q(s, o)$ by choosing in each state s the option o that maximizes its value. The termination condition $\beta(s)$ is set to 1 for $s = \overline{s}^k$ and to 0 elsewhere. For what concerns the initial set, it can be limited to a subset of the state space $\mathcal{S}$ composed by the states that have been most visited in the skill acquisition phase.

The incremental generation of new options makes the agent able to develop a hierarchy of skills, where new options can reuse previously learned options to achieve the goals extracted in the motivating phase.

4 Experiments

In this section, we provide experimental results obtained by SMILe in two different environments. In the first problem, we show how SMILe learns general purpose options that may be effectively reused for learning to reach a large number of goals. The second experiment puts in evidence how the SMILe development process can significantly reduce the learning times.

4.1 Four-Room Gridworld

The *Four-Room* (Fig. 2) environment [18] is a 10x10 grid with a set of walls that delimit four rooms. The initial set of actions is $\mathcal{A} = \{down, right, up, left\}$

and the starting state is the upper left corner. To introduce stochasticity in the world dynamics, actions have a probability of 0.3 to fail. When an action fails the agent moves to one of the adjacent states at random.

The development process of SMILe led to the identification of interesting goals only in limited regions. The density plot in Fig. 3 shows the frequency of subgoal identification for each state: the lighter the region the higher the frequency of extraction. It is worth noting that, using the interest function described in Section 3, SMILe finds the states in the middle of the rooms as most interesting. Recalling the definition of local interest (Eq. 1), the explanation of this result is that from these states the agent can easily reach all the other states in the room. The usefulness of the learned skills can be measured only by imposing many different external goals to the agent and by evaluating the global learning performance. Therefore, we have performed a comparison among Q-learning [17], Q-learning with four skills whose goals have been chosen at random, and Q-learning with the four skills learned by SMILe, over 1000 randomly extracted external goal states. Then, we have recorded the sum of learning steps for each goal over the first 100 episodes. Table 1 reports the number of steps in the average and in the worst case. As it can be noticed, both the average and the maximum number of steps needed by SMILe are less than those needed by the other two algorithms. This means that the skills acquired by SMILe produced a relevant advantage when facing different learning problems. Furthermore, since Q-Learning with skills for random goals obtained the worst performance, the result of SMILe is not simply determined by the use of the option framework, but it strongly depends on the identification of interesting states that lead to the acquisition of general-purpose skills.

4.2 Playworld Environment

The second experiment we discuss, is a version of the Playworld proposed in [16]. The Playworld is an abstraction of a real environment characterized by two rooms with a door in between, two panels and a charger (see Fig. 4). The panels are in the room at left: the light panel switches the light on and off, while the door panel opens and closes the door. The animat perceives the light intensity, whether the door is open or not, its charge level and its position (i.e., absolute coordinates and orientation). The animat is initially placed at random in the left room and the light is switched off. When in the dark, the animat may fail in taking the selected action with a probability of 0.2, it cannot perceive the status of the door, and the door panel is deactivated. Once the light is switched on, actions always succeed, the animat can open the door, move to the other room and charge. The animat can turn left, turn right, and move ahead.

The experiment consists of two main stages: intrinsically motivated incremental learning and extrinsically motivated learning. In the first stage the animat explores the environment and develops new skills according to the process described in Section 2. In the second stage, five different goals are imposed by an external designer by providing an extrinsic reward function.

In the first stage, the salient events we can expect the animat to find are: *light on*, *light off*, *open door*, *close door*, *charge*. The upper graph of Fig. 5 shows the events occurred in the babbling phase at first iteration, when the agent succeeds in switching the light on and off only a few times. The lower graph of Fig. 5 shows the changes in the babbling phase introduced by the skills learned after five iterations. As it can be noticed, the skills developed in the previous iterations bias the random exploration so that the animat succeeds in activating new and more complex events (e.g., open the door and charge). This shows how SMILe enables the animat to autonomously discover interesting configurations in the environment and to develop new skills for achieving them.

In the second stage, when the development process is over, we compare the performance of an animat that exploits the new skills, to that of an animat using Q-Learning with basic skills, on five different tasks:

Task1: charge
Task2: charge, move to upper left corner of right room
Task3: charge, move to upper left corner of left room
Task4: charge, move to left room and close the door
Task5: charge, move to left room, close the door, switch the light off

While *Task2* and *Task3* are not strictly related to any salient event, the other tasks require the animat to achieve configurations relevant in the Playworld environment. In the comparison, we adopted the same learning parameters for both Q-Learning and SMILe (learning rate $\alpha = 0.6$, ϵ-greedy exploration with $\epsilon = 0.2$, discount factor $\gamma = 0.95$). Each 1,000 learning episodes, the extrinsic reward function is changed according to the task that must be accomplished and the learning animat should be able to adapt its policy to the new task without restarting the learning from scratch.

Fig. 6 shows the number of steps per learning episode. The first 2,100 episodes, labeled as *Self-Development* in the graph, represent the first stage of the experiment in which the SMILe animat autonomously identifies six different interesting states, used as goals for learning six new skills. On the other hand, in the first stage the Q-Learning animat does nothing, since no extrinsic reward is provided. The second stage starts with the introduction of a positive extrinsic reward for achieving the charger. While the Q-Learning animat can only use the basic skills, the SMILe animat exploits the skills learned in the first stage and succeeds in finding the optimal policy to reach the charger in less episodes than those needed by Q-Learning. Similarly, SMILe succeeds in exploiting its skills even for changing tasks, while Q-Learning took more time to adapt to new extrinsic reward functions. Furthermore, in Fig. 7 we compare the total number of steps for both the algorithms and we report their difference. In the first stage, SMILe takes almost 250,000 steps to explore the environment and to learn the new skills, while no steps are taken by the Q-Learning robot. Notwithstanding the initial loss, the total number of steps needed by SMILe after the accomplishment of *Task1* is less than that of Q-Learning. The advantage of SMILe becomes even more relevant at the end of the second stage when Q-Learning took almost twice as many steps as SMILe. This comparison shows that SMILe, even though it requires po-

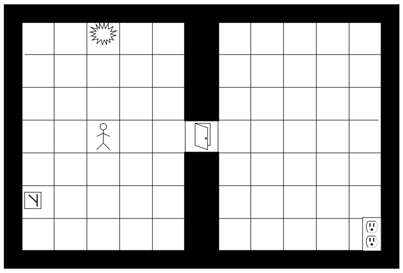

Fig. 4. The Playworld environment

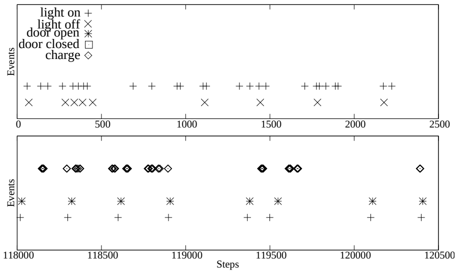

Fig. 5. Sequence of events in the first (*upper*) and fifth (*lower*) babbling phase.

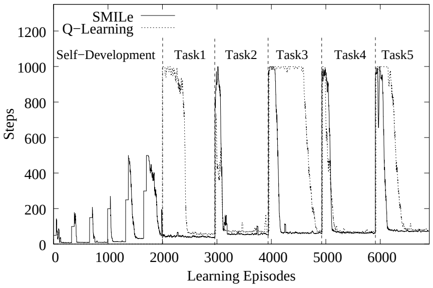

Fig. 6. Comparison of performance between Q-Learning and SMILe.

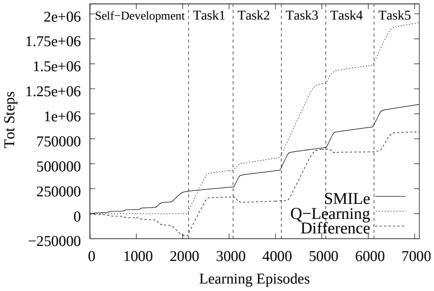

Fig. 7. Comparison of total number of steps between Q-Learning and SMILe.

tentially expensive exploration of the environment, leads to the development of useful skills that can be profitably reused in many different tasks. In particular, the number of steps saved during the extrinsically motivated learning stage is greater than those used in the first stage already in the first goal.

5 Conclusions

Hand-coded abilities, though useful in domains where tasks are fixed, proved to be inadequate to enable artificial systems to solve even slightly different tasks in uncertain environments. On the other hand, the capability to develop new skills from basic abilities without any imposed goal, is what makes human beings and animals able to reuse their skills in many complex tasks.

In this paper, we have presented SMILe, a self-development RL framework that incrementally acquires more and more complex skills through an iterative three phase learning process similar to those taken by children and animal puppies in their early development stages. Experimental results show the effectiveness of the skills learned by SMILe when operating in environments where different tasks may arise, thus developing agents with a good degree of autonomy.

Currently, we are investigating the use of function approximation techniques to scale to large, high dimensional domains. Future work includes the integration of SMILe with developmental robotics approaches in real robotic tasks [6].

References

1. A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(4):341–379, 2003.
2. A.G. Barto, S. Singh, and N. Chentanez. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of ICDL*, 2004.
3. D. E. Berlyne. *Conflict, Arousal, and Curiosity.* McGraw-Hill, 1960.
4. A. Bonarini, A. Lazaric, and M. Restelli. Smile: Self-motivated incremental learning. Technical report, Politecnico di Milano, http://www.airlab.elet.polimi.it/papers/bonarini06smile.pdf, 2006.
5. S. Kakade and P. Dayan. Dopamine: Generalization and bonuses. *Neural Networks*, 15:549–559, 2002.
6. G.D. Konidaris and G.M. Hayes. An architecture for behavior-based reinforcement learning. *Adaptive Behavior*, 13(1):5–32, 2005.
7. M. Lungarella, G. Metta, R. Pfeifer, and C. Sandini. Developmental robotics: a survey. *Connection Science*, 15(4):151–190, 2003.
8. J. Marshall, D. Blank, and L. Meeden. An emergent framework for self-motivation in developmental robotics. In *Proceedings of ICDL*, 2004.
9. A. McGovern and A. G. Barto. Automatic discovery of subgoals in reinforcement learning using diverse density. In *Proceedings of ICML*, 2001.
10. A. Meltzoff and M. Moore. Explaining facial imitation: a theoretical model. *Early Development and Parenting*, 6:179–192, 1997.
11. I. Menache, S. Mannor, and N. Shimkin. Q-cut - dynamic discovery of sub-goals in reinforcement learning. In *Proceedings of ECML*, 2002.
12. P-Y. Oudeyer, F. Kaplan, V. Hafner, and Whyte A. The playground experiment: Task-independent development of a curious robot. In *AAAI Spring Symposium Workshop on Developmental Robotics*, 2005.
13. J. Piaget. *The Origins of Intelligence in Children.* Norton, N.Y., 1952.
14. B. Ratitch and D. Precup. Using mdp characteristics to guide exploration in reinforcement learning. In *European Conference on Reinforcement Learning*, 2003.
15. J. Schmidhuber. Self-motivated development through rewards for predictor errors / improvements. In *AAAI Spring Symposium on Developmental Robotics*, 2005.
16. A. Stout, G. Konidaris, and A. Barto. Intrinsically motivated reinforcement learning: A promising framework for developmental robot learning. In *AAAI Spring Symposium on Developmental Robotics*, 2005.
17. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.
18. R. S. Sutton, D. Precup, and S.P. Singh. Between mdps and semi-mdps: a framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999.
19. E. Uchibe and K. Doya. Reinforcement learning with multiple heterogeneous modules: A framework for developmental robot learning. In *Proceedings of ICDL*, 2005.
20. J. Weng, A. McClelland, O. Sporns, I. Stockman, M. Sur, and E. Thelen. Autonomous mental development by robots and animals. *Science*, 291:599–600, 2001.
21. J. Weng and Y. Zhang. Novelty and reinforcement learning in the value system of developmental robots. In *International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, 2002.