Selective Exploration Exploiting Skills in Hierarchical Reinforcement Learning Framework

Gakuto Masuyama[1], Atsushi Yamashita[2] and Hajime Asama[2]

Abstract— **In this paper, novel reinforcement learning method with intrinsic motivation for reproducibility of the past successful experience is presented. The experience is extracted as skill, which is composed of action sequence and abstract knowledge about observed sensor input. Utilizing the collected skills, reproduction of the successful experience is attempted in novel and unknown environment. Consistent exploration and active reduction of search space are realized by learning with intrinsic motivation for reproducibility of experience. Simulation experiments in grid world demonstrate that proposed method significantly accelerate speed of learning.**

I. INTRODUCTION

Development of autonomous robot that learns incrementally without sufficient prior knowledge is a challenging open problem in robotics research. Reinforcement learning [1] has been actively studied to tackle with the problem, because it does not require precise description about task, environment, and dynamics of robot. In recent years, introduction of intrinsic motivation to hierarchical reinforcement learning [2], [3] is receiving much attention in the context of developmental robotics.

Traditional reinforcement learning methods are typically designed for single, isolated task. Therefore they must start learning from scratch for every newly given task, and it is difficult to achieve sufficient learning speed as control architecture for autonomous and adaptive robot. In contrast, animals, including human, can demonstrate adaptability to novel tasks and environments instantaneously. One of potent factors of such a difference seems to be mechanism of intrinsic motivation. Psychologists have separated intrinsic motivation from extrinsic motivation [4]. Extrinsic motivation drives organisms to take purposive behavior. On the other hand, purpose of behavior driven by intrinsic motivation is behavior itself, *e.g.* behavior driven by curiosity or interest. Intrinsically motivated activities are explorative. The motivational system is not designed to deal with specific single task; it is designed to acquire increasing competence.

A number of studies have been presented that show effectiveness of intrinsic motivation implemented on reinforcement learning. Recently intrinsic motivation is utilized for autonomous acquisition of useful skills in hierarchical reinforcement learning framework. Skill (also called option, temporally extended action, etc.) is a multi-step partial policy that enables learner to streamline decision making process. If the agent can acquire skills that are applicable to various situations, it should be possible to accelerate learning of current task. In [5], intrinsic reward for salient event to acquire useful skills are applied to hierarchical reinforcement learning framework. The agent learns skills incrementally based on saliency in the environment, and utilize skills to improve learning performance. Consequently, the agent acquires hierarchical structure of skill in response to experienced sensory-motor flow.

These studies do not necessarily optimize given a whole problem, however, skill is extracted from optimal policy for a sub-task (*cf.* recursively optimal policy [6]). That is, skill is a map from subspace of state space to action space. If skill is formulated as such a closed partial policy, collection of exhaustive experience is required. Although the skill is defined in smaller space than state space in general, the formulation could lead to deceleration of learning. Maturation of learner would increase complexity of skills, and accordingly enlargement of skill learning process would occur apparent stagnation of learning. Der pointed out that such a problem is a kind of curse of dimensionality [7].

Pursuit of optimality for sub-task is fundamental for autonomous robot to acquire broad competence, but at the same time, autonomous robot must have an ability to adapt novel situation instantaneously. In this perspective, it is rational to find a few tentative solutions at first, and after that, learn all-inclusive optimal solution. On the basis of above methodology, reinforcement learning framework with intrinsic motivation for reproduction of the past successful experience is presented in this paper. A key idea is self-motivated reduction of search space using skill, which represents the past successful experience. Contrast to previous studies that determine exploration strategy based on maximization of learning progress [8], [9], selective exploration is implemented via exploitation using skills. Proposed method does not guarantee global optimality in general, however, learning speed is significantly accelerated when the agent finds sub-optimal solutions.

A. Overview

Fig. 1 depicts framework of proposed method. Proposed framework has hierarchical structure that is composed of one-step primitive action (hereinafter simply called "action") and skill that assigns multi-step action series. Different update laws are applied to learn action value and skill value unlike option framework [10]. Q-learning [11] is applied to

[1]G. Masuyama is with the Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 113-8551, Japan. masuyama at mech.chuo-u.ac.jp

[2]A. Yamashita and H. Asama are with the Department of Precision Engineering, Faculty of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan. {yamashita, asama} at robot.t.u-tokyo.ac.jp

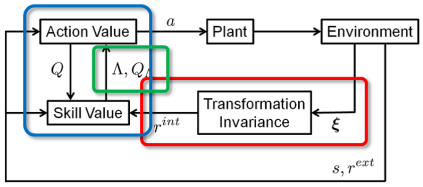

Fig. 1. Framework of proposed method

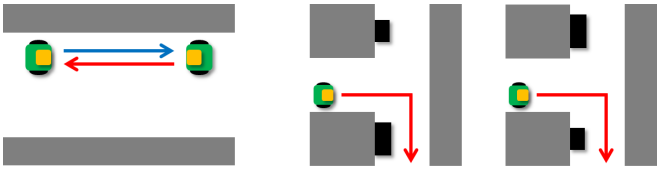

Fig. 2. Identifiable experiences accompanied by skill execution

learn action value. Skill value is learned in Semi-Markov Decision Processes (SMDPs) using three elements: intrinsic reward signal for reproducibility of the past successful experience, extrinsic reward signal derived from given task, and action value. Hence state-skill pairs that result in

- observation of experience with high reproducibility,
- action selections toward task completion,
- transition to a good state (high state value),

take high skill value.

Each state-skill pair would have almost uniform value in early phase of learning if prior information is not given. Then update of skill value strongly commits intrinsic reward signal at first. If the executed skill generates high reproducibility of the past successful experience, positive intrinsic reward is given to the skill, and consequently the skill would take relatively high value. This is one of a key point of proposed method, because skill value has a role to characterize exploration strategy. In each action selection step, action value is instantaneously biased by value of currently activating skill. If the value of activating skill is positive, then probability to select an action assigned by the skill is increased. Thus search space is reduced by skills that have high value due to intrinsic reward signal, especially in early phase of learning. Action value (skill value) is not learned for every state-action (state-skill) pair; the agent explores and learns selectively using skills. Acceleration of learning speed can be realized due to the active reduction of search space.

In Section II, affine transformation invariant feature is introduced to represent experiences accompanied by skill execution. It enables the agent to evaluate experience of skill execution in a novel environment abstractly. Update law of skill value and instantaneous bias to action selection process are described in Section III. Simulation experiments supposing navigation problem of mobile robot is conducted in Section IV. Proposed method and one of a representative hierarchical reinforcement method [10] are simulated. Additionally, fundamental property of proposed framework is discussed. Finally, conclusions are summarized in Section V. Short abstract of proposed framework is presented in [12]. In this paper, practical algorithms, thorough results, and design concepts are detailed. Comparison with option framework [10] is also presented.

Note that proposed method differs from knowledge transfer [13] such as initialization of value function. Skills are used to learn sub-optimal path rather than optimal map in this paper. Additionally, incremental skill acquisition during learning is not considered explicitly. It would helpful to improve performance of learning, however, validation for active reduction of search space is a focus of this paper. Therefore temporal development of the learner is excluded, and the learner is supposed to have determinate skills. Integration with autonomous skill acquisition processes is discussed in a future work.

II. ABSTRACTION OF EXPERIENCES

Abstract representation for a sequence of sensor inputs, which is associated with skill execution, is introduced to generate intrinsic reward signal. There could be various measures for reproducibility of the past successful experience. Cyclopedic collection and evaluation for every possible experience are not realistic. Distance between two sequences might be applicable, however, occasionally it would be irrational. As a simple illustration, let us consider wall following behavior of a mobile robot equipping range sensor. For a robot moving along wall, it would not be worth distinguishing "wall existing side" to a moving direction. Distance from the wall should be sufficient information as long as the robot just continues to follow the wall. In other words, it is reasonable to identify symmetric information about axis of translational movement in the case as is shown in Fig. 2. Similar discussion would be also valid for more general cases, such as right turn at T-junction, and passing through with pedestrians. Therefore abstraction of limited experience is an important issue to reduce search space.

To abstract a sequence of sensor inputs accompanied by skill execution, affine transformation invariant feature [14] $M \in \mathbb{R}$ is introduced in proposed method. Affine transformation invariant feature is developed in speech recognition research domain. Let $\boldsymbol{\xi}_t \in \mathbb{R}^d$ be feature vector of sensor input, and let $\boldsymbol{\Xi}_{t-k_1:t+k_2} = [\boldsymbol{\xi}_{t-k_1}, ..., \boldsymbol{\xi}_t, ..., \boldsymbol{\xi}_{t+k_2}]$ be a sequence of $\boldsymbol{\xi}$. Let $\bar{\boldsymbol{\xi}}_t = \boldsymbol{A}\boldsymbol{\xi}_t + \boldsymbol{c}$ be affine transformation for $\boldsymbol{\xi}_t$, then affine transformation invariant feature M for $\boldsymbol{\Xi}_{t-k_1:t+k_2}$ satisfies $M(\boldsymbol{\Xi}_{t-k_1:t+k_2}) = M(\bar{\boldsymbol{\Xi}}_{t-k_1:t+k_2})$, where $\bar{\boldsymbol{\Xi}}_{t-k_1:t+k_2} = [\bar{\boldsymbol{\xi}}_{t-k_1}, ..., \bar{\boldsymbol{\xi}}_t, ..., \bar{\boldsymbol{\xi}}_{t+k_2}]$. An actual functional form of the invariant feature is as below

$$M(\boldsymbol{\Xi}_{t-k_1:t+k_2}) = \sqrt{(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^T(\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b)^{-1}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)}. \tag{1}$$

$\boldsymbol{\mu}_a$ and $\boldsymbol{\Sigma}_a$ are mean and covariance matrix of $\boldsymbol{\Xi}_a := \boldsymbol{\Xi}_{t-k_1:t-1}$ respectively, which is an arbitrary subsequence of $\boldsymbol{\Xi}_{t-k_1:t+k_2}$.

$$\boldsymbol{\mu}_a = \frac{1}{k_1} \sum_{\tau=t-k_1}^{t-1} \boldsymbol{\xi}_\tau, \tag{2}$$

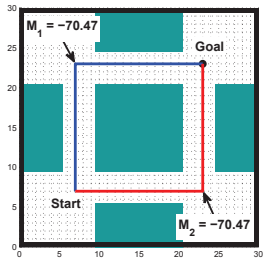

Fig. 3. Symmetric trajectories

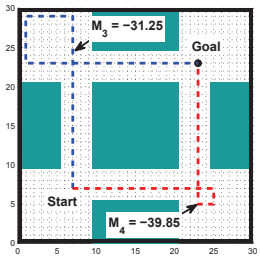

Fig. 4. Asymmetric trajectories

$$\mathbf{\Sigma}_a = \frac{1}{k_1} \sum_{\tau=t-k_1}^{t-1} (\boldsymbol{\xi}_\tau - \boldsymbol{\mu}_a)(\boldsymbol{\xi}_\tau - \boldsymbol{\mu}_a)^T. \tag{3}$$

$\boldsymbol{\mu}_b$ and $\mathbf{\Sigma}_b$ are also defined for subsequence $\mathbf{\Xi}_b := \mathbf{\Xi}_{t:t+k_2}$. It is well known that the difference of vocal tract length and recording equipment can be approximately modeled by affine transformation for cepstral vectors. Thus affine transformation invariant feature offers a measure of inherent structure in particular speech language without normalization of massive data.

Similar consideration could be valid for arbitrary sequences of sensor inputs observed through skill execution. As a consequence of interaction between an agent and an environment over time, a series of sensor inputs is observed. Each skill execution segments the series. The segmented sensor inputs forms geometric structure in feature space. For the structure, transformation invariance removes unnecessary information modeled by certain transformation. Thus it can be utilized to measure the unique structure in feature space (*e.g.* sequence of phoneme in speech recognition) that corresponds to certain action sequence (utterance of a word). There are studies utilizing invariant property for bootstrapping [15], [16], but they are different from our idea that we focus on temporally extended data sequence. In this paper, utility of such an idea is verified through sensor settings supposing mobile robot, which directly reflect geometric structure of an external environment.

Fig. 3 and Fig. 4 illustrate examples of observed affine transformation invariant feature in two dimensional grid world. Obstacles (depicted as green box in the figures) are set in the environments. An agent can select four actions: move one cell to the up, right, down, left directions. The agent is associated to equip odometer and range sensor. Odometer measures distance from current state to initial state, and range sensor measures distance from current state to state of obstacles with respect to four directions: up, right, down, and left. Feature vector $\boldsymbol{\xi} = [\xi_1; \xi_2]$ is chosen as 2 dimensional vector: ξ_1 is distance from initial state, and ξ_2 is sum of range sensor readings. With these settings, transformation invariant feature was calculated for determined sequence of actions.

In Fig. 3, sequences of ξ_1 is the same between the two paths (red and blue lines), but sequences of ξ_2 are different. However, the same affine invariant feature values $M_1 = M_2 = -70.47$ are observed for each path. This result means that sequence of $\boldsymbol{\xi}$ in one path can be identified with $\boldsymbol{\xi}$ in another path by a certain affine transformation. In Fig. 4, two paths are not symmetric, but have similar "walk around " movement in the middle of traveling. Observed invariant features are $M_3 = -31.25$ for longer path (blue dashed line) and $M_4 = -39.85$ for shorter path (red dashed line), respectively. Therefore the two experiences are discriminated. Considering invariant feature in Fig. 3, $|M_1 - M_3| > |M_1 - M_4|$. Thus, in the sense of affine invariant feature, shorter path in Fig. 4 is more similar with no walk around path than the longer path in Fig. 4. On the other hand, $|M_1 - M_4| < |M_3 - M_4|$. Thus existence of walk around movement makes greater distinction than the difference of length of walk around movement. These results exemplify the function of affine invariant feature as the measure of sequence of feature vector derived from certain action series, or skill. That is to say, the agent can measure the degree of similarity in two experiences from M value [1].

III. SELECTIVE EXPLORATION AND LEARNING

In this section proposed learning method is detailed. Here S and A denote state space and action space, and $s \in S$ and $a \in A$ denote a state and an action respectively. Skill is formally defined as $\Lambda = (A_\Lambda, M_\Lambda)$. $A_\Lambda = \{a_\Lambda(1), a_\Lambda(2), \cdots, a_\Lambda(T_\Lambda)\}$ is a finite ordered action set and $M_\Lambda \in \mathbb{R}$ is an affine invariant feature observed by execution of A_Λ. A_Λ and M_Λ are supposed to be acquired in successfully solved sub-task in the past. Action value $Q(s, a)$ is learned by Q-learning [11] in proposed method. In parallel with Q-learning, skill value $Q_\Lambda(s, \Lambda)$ is learned with intrinsic reward signal in SMDPs framework (see Fig. 1). Skill value is a function that represents a value to select skill Λ in state s. Action value is instantaneously biased by value of currently activating skill in every action selection process. Therefore explorative behavior of the agent is characterized by skills and their value.

A. Update Law of Action and Skill Value

Action value is updated by ordinal Q-learning:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha \left(r_{t+1}^{ext} + \gamma \max_{a'} Q(s', a') \right). \tag{4}$$

$\alpha \in [0, 1]$ is learning rate and $\gamma \in [0, 1]$ is discount rate. $r_{t+1}^{ext} \in \mathbb{R}$ denotes immediate extrinsic reward and $s' \in S$ is state at next time step after execution of action a. Update process of action value is not affected explicitly by skill value. Therefore arbitrary action selection method for Q-learning is applicable such as ε-greedy and soft-max action selection [1].

[1]For more complex sensor settings, appropriate preprocessing of sensor information would be required [14]. Additionally, we need to consider carefully what kind of property, i.e. transformation should be ignored. However, these topics relate characteristics of specific sensor settings and they are out of scope of this paper. Therefore we associate simple sensor settings and show intuitively interpretable results.

Update law of skill value $Q_\Lambda(s,\Lambda)$ is similar with update law of Q-learning applied to SMDPs [2]:

$$Q_\Lambda(s_\Lambda,\Lambda) \leftarrow (1-\alpha)Q_\Lambda(s_\Lambda,\Lambda) + \alpha\left(r^{int} + R_\Lambda + \gamma^{T_\Lambda}\max_{a''} Q(s'',a'')\right). \quad (5)$$

It is updated only when action sequence of the skill is completed or the next state is the destination state. $s_\Lambda \in S$ denotes state where skill has started and s'' denotes state after execution of Λ. $R_\Lambda \in \mathbb{R}$ is a discounted cumulative extrinsic reward gained during skill execution. R_Λ is calculated as

$$R_\Lambda = \sum_{t_\Lambda=1}^{T_\Lambda} \gamma^{t_\Lambda-1} r^{ext}_{t+t_\Lambda}. \quad (6)$$

As natural formulation for temporal extension of action, immediate extrinsic reward during skill execution is discounted and accumulated, and finally utilized to update skill value.

Intrinsic reward $r^{int} : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is calculated by below equation using affine transformation invariant feature as a measure of reproducibility of experience:

$$r^{int}(M, M_\Lambda) = r_p^{int}\exp\left(-\frac{|M-M_\Lambda|}{\rho}\right). \quad (7)$$

M is observed affine invariant feature through execution of Λ in present environment. M_Λ denotes affine invariant feature of Λ. $r_p^{int} \in \mathbb{R}_+$ and $\rho \in \mathbb{R}_+$ denote positive parameters. Thus skill that demonstrates high reproducibility in the sense of transformation invariant feature tends to be reinforced positively. Note that the intrinsic reward is independent of given task. Therefore intrinsic reward itself does not assure improvement of performance.

B. Bias for Action Selection Process by Skill Value

Action value is utilized for update law of skill value in (5). On the other hand, skill value is utilized for action selection. Action value is biased by skill value at each time step, and skill is embedded to action value implicitly. In the course of skill execution through time, a value of action $a_\Lambda(t_\Lambda)$, which is assigned by executing skill, is modified in accordance with the skill value. The instantaneous bias for relevant action value is calculated as

$$Q(s,a_\Lambda) \xleftarrow{bias} Q(s,a_\Lambda) + \beta\gamma^{-(t_\Lambda-1)}Q_\Lambda(s_\Lambda,\Lambda). \quad (8)$$

t_Λ denotes elapsed time from start of execution of Λ. $a_\Lambda \in A$ is action assigned by skill at time t_Λ, and $s_\Lambda \in S$ is a starting state of skill execution. $\beta \in \mathbb{R}$ is a parameter determining intensity of bias. β is supposed to be positive in this paper. Note that (8) is not an update law; the bias is one-time adding. Therefore the value of $Q(s,a_\Lambda)$ is turned back once the action is selected.

Fig. 5 depicts conceptual scheme of relationship between $Q(s,a)$ and $Q_\Lambda(s,\Lambda)$. Relevant action to activating skill is biased by the value of skill. If the value of activating skill is positive (negative), then probability to select relevant action is increased (decreased). In other word, exploration strategy of agent is characterized by skills. Entire procedure of proposed method is summarized in Algorithm 1.

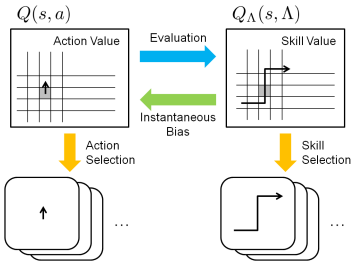

Fig. 5. Relationship between action value $Q(s,a)$ and skill value $Q_\Lambda(s,\Lambda)$

Algorithm 1 Summary of algorithm

```
initialize Q(s,a), Q_Λ(s,Λ)
repeat
  initialize s
  select skill Λ using Q_Λ(s,Λ)
  repeat
    bias Q(s,a_Λ) by (8)
    select a using biased Q(s,a)
    get sensor data and calculate ξ
    observe r^ext_{t+1} and next state s'
    update Q(s,a) and R by (4), (6)
    if t_Λ = T_Λ then
      calculate M and r^int by (1) and (7)
      update Q_Λ(s_Λ,Λ) by (5)
      select next skill Λ' using Q_Λ(s',Λ)
    end if
    t ← t+1, t_Λ ← t_Λ+1, s ← s'
  until termination condition of episode is satisfied
  update Q_Λ(s_Λ,Λ) by (5)
until termination condition of learning is satisfied
```

IV. SIMULATION EXPERIMENT

To verify validity of proposed method, simulation experiments for navigation tasks in two dimensional grid world is demonstrated. Fig. 6 depicts an example of tested environment. A circle in the figure represents initial state, and a star represents destination state. Fig. 7 depicts an example of skill acquisition environment and acquired skill. Skill is supposed to be acquired from execution of optimal policy for sub-task in this paper. Executed sequence of actions and resulted affine transformation invariant feature is stored as skill. Results of proposed method, option framework [10], and Q-learning are shown.

A. Simulation Settings

The agent is available four actions: moving one cell to the up, right, down, and left directions. Tested environments are square walled rooms (Fig. 6). There are multiple obstacles that occupy one cell, but obstacles do not exist at cells bounding the wall. Positions of obstacles are randomly determined at the beginning of learning. Initial state of the agent is $(1,1)$ and destination state is $s_d = (15,15)$. Extrinsic reward is $+5$ at destination state, -1 for collision with obstacles, and -0.1 for each action execution. Initially $Q(s,a) = 0$ for all state-action pairs. $Q_\Lambda(s,\Lambda) = 1$ for all state-skill pairs to

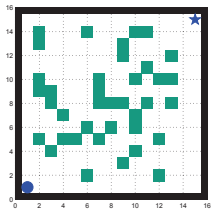

Fig. 6. Example of task

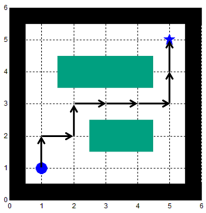

Fig. 7. Environment for skill setting

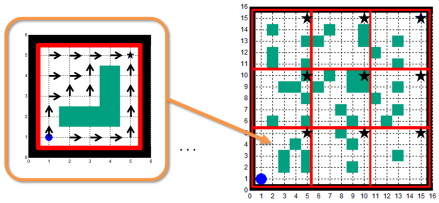

Fig. 8. Settings of option

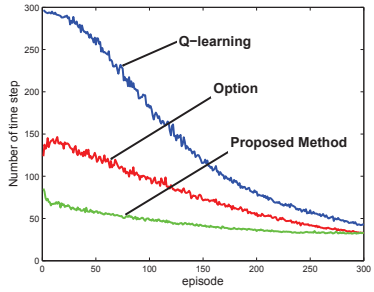

Fig. 9. Elapsed time step

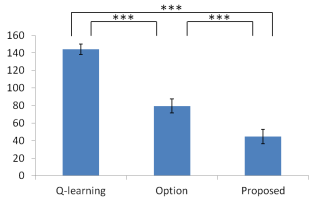

Fig. 10. Mean of elapsed time step

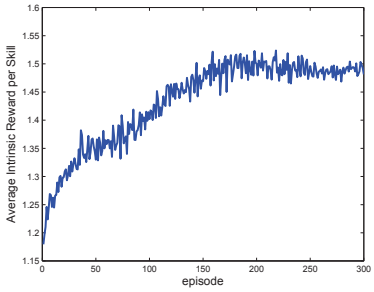

Fig. 11. Intrinsic reward per one skill execution

bias action value. $\rho = 10$, $r_p^{int} = 2$, $\alpha = 0.2$, $\gamma = 0.95$, $\beta = 0.5$, and termination conditions of episode are reaching s_d or elapse of 300 steps. Action and skill selection policy is soft-max action selection [1]. Temperature parameter $\tau \in \mathbb{R}$ is set as $\tau = 0.2\exp(-e/200)$, here e is current number of episode. The agent equips range sensors that detect over two cells for four directions and an odometer. Feature vector $\boldsymbol{\xi} = [\xi_1; \xi_2]$ is chosen as 2 dimensional vector; ξ_1 is distance from initial state and ξ_2 is sum of range sensor readings. Sequence of $\boldsymbol{\xi}$ is segmented in the half point of the skill to calculate affine invariant feature.

Skill is supposed to be acquired preliminary in an environment where $s_d = (5, 5)$ and randomly positioned five obstacles are set (Fig. 7). Q-learning is implemented for the sub navigation task. Executing learned greedy policy, a sequence of actions and M are memorized as a skill. The environment is initialized and above procedure is repeated five times (i.e. five skills are acquired).

In option framework, *interruption* and *intra-option learning* are applied [10]. Q-learning is implemented and five optimal policies are learned in the same way as the skill setting processes. Then test environments are segmented into nine areas, and the five optimal policies are assigned to every area as is shown in Fig. 8. Option policy is deterministic (greedy). Randomly selected actions are assigned to states corresponding to obstacles in option acquisition environments.

B. Results

With above settings, 30 trials were implemented. Obstacle, skill, and option settings are initialized for every trial. For one trial, mean of 10 experiments was used for results. The agent learned 300 episodes in one experiment.

Fig. 9 depicts transition of elapsed time step for each episode. Vertical axis represents elapsed time step in one episode, and horizontal axis represents episode. In Fig. 10, average time step of 30 trials with respect to 300 episodes is shown. Paired t-test is performed, and the significant difference was found between every two methods ($p < 0.001$). It can be seen that proposed method converged significantly faster than Q-learning, although update law of two methods are the same. State-action space is explored isotropically centering around initial state in Q-learning. Therefore Q-learning can assure optimality of learned policy under certain conditions, though it requires exhaustive exploration. On the other hand, proposed method does not explore entire state-action space. The agent explores in accordance with skills, which have certain consistency as successful experiences in the past. Utilization of skills might not help to find positive extrinsic reward in early phase of learning, but the agent makes intrinsic reward for reproducibility of the past successful experience. As a result, some paths are composed that fill a role of center for exploration. If one of the paths reaches positive extrinsic reward, then search space is reduced furthermore. Once rough path to the destination state is discovered, update of action value makes smaller refinement.

Fig. 11 shows transition of intrinsic reward. Vertical axis represents received intrinsic reward per one skill execution,

and horizontal axis represents episode. Intrinsic reward, or reproducibility of experience increased in accordance with progress of learning. This result shows that explorative behavior was controlled by intrinsic reward so as to increase the reproducibility. Stagnation from around 200 episode is due to the refinement of path by action value mentioned above. Finally converged solution means that influence of action and skill value balanced. It is remarkable that skills demonstrating high reproducibility do not necessarily have a priority to be executed. Available skills may not be perfectly appropriate for the task. Although sometimes execution of skill would significantly degenerate performance, it can be helpful to accelerate learning. Because ill-fitting skills would provide negative biases to the agent, they contribute to reduce search space.

Proposed method also converged faster than option framework (Fig. 9, Fig. 10). Options have the same consistency as skills, because option policies are optimal policy for subproblems. Option framework biases exploration in action space, but it does not bias exploration in state-action space. This is significant difference between proposed method and hierarchical reinforcement learning methods that take their stand on concept of optimality. The results indicate effectiveness of selective exploration using abstract knowledge about the past successful experiences.

C. Discussion

Fundamental idea of proposed framework is active reduction of search space using skills, i.e. knowledge. The idea is inspired by top-down process in information process system of humans [17]. Top-down process generates "expectation" for existence of an event that is relevant to certain knowledge, and it biases information process system to recognize expected results. Thus the bias works to measure current experience using knowledge, which is acquired in the past. Such a function of subjective conceptualization should be required for autonomous robot that needs to adapt unforeseen situations instantaneously.

In this paper, acceleration of learning speed is shown for navigation task of abstracted mobile robot, however, there are important open problems. Proposed method supposes discrete state-action space, but it is well known such a formalization has curse of dimensionality problem. Therefore proposed framework should be extended to continuous state-action space domain using function approximation techniques. Additionally, dynamics of robot is ignored in this paper. When dynamics of robot is taken into account, context of skill execution should be considered. Integration with methods that autonomously construct hierarchical structure of skills, which are actively studied [5], [9], would be promising on this point.

V. CONCLUSIONS

In this paper, novel intrinsically motivated reinforcement learning method is presented. Selective exploration using skill, which is sequence of primitive actions and abstract knowledge observed in the past successful experience, is a key of the method. The abstract knowledge is geometric transformation invariant feature calculated from sequence of sensor readings. Reproducibility of the past successful experience is measured in a novel environment using the invariant feature. With that, highly reproducible skill is intrinsically motivated. Action value and skill value are learned in parallel, and action value is instantaneously biased by skill value in action selection process. As a result, consistently extended actions, or skill is embedded to action value adaptively. It is verified through simulation experiments that proposed method realize selective exploration, and significantly accelerate learning progress.

Integration with incremental skill acquisition process is future work. Authors are considering to apply proposed framework to model-based reinforcement learning method in parallel with the above integration.

REFERENCES

[1] R.S. Sutton, A.G. Barto, Reinforcement Learning:An Introduction, Cambridge, MA, MIT Press, 1998.

[2] A.G. Barto, S. Mahadevan, Recent Advances in Hierarchical Reinforcement Learning, , Discrete Event Dynamical Systems: Theory and Applications, vol.13, pp.341–379, 2003.

[3] A.G. Barto, Intrinsic Motivation and Reinforcement Learning, Intrinsically Motivated Learning in Natural and Artificial Systems, Springer Berlin Heidelberg, pp.17–47, 2013.

[4] P.-Y. Oudeyer, F. Kaplan, How Can We Define Intrinsic Motivation?, Proc. of the 8th Int'l Conf. on Epigenetic Robotics, 2008.

[5] S. Singh, A.G. Barto, N. Chentanez, Intrinsically Motivated Reinforcement Learning, Proc. of Advances in Neural Information Processing Systems 17, pp.1281–1288, 2005.

[6] T.G. Dietterich, Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition, J. of Artificial Intelligence, vol.13, pp.227–303, 2000.

[7] R. Der, G. Martius, From Motor Babbling to Purposive Actions: Emerging Self-exploration in a Dynamical Systems Approach to Early Robot Development, From Animals to Animats 9, Lecture Notes in Computer Science, Springer Berlin/Heidelberg, vol.4095, pp.406–421, 2006.

[8] P.-Y. Oudeyer, F. Kaplan, V.V. Hafner, Intrinsic Motivation Systems for Autonomous Mental Development, IEEE Trans. on Evolutionary Computation, vol.11, no.2, pp.265–286, 2007.

[9] C.M. Vigorito, A.G. Barto, Intrinsically Motivated Hierarchical Skill Learning in Structured Environments, IEEE Trans. on Autonomous Mental Development, vol.2, no.2, pp.83–90, 2010.

[10] R.S. Sutton, D. Percup, S. Singh, Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning, Artificial Intelligence, vol.112, pp.181–211, 1999.

[11] C.J.C.H. Watkins, P. Dayan, Q-learning, Machine Learning, vol.8, pp.279–292, 1992.

[12] G. Masuyama, A. Yamashita, H. Asama Intrinsically Motivated Anticipatory Learning Utilizing Transformation Invariance, Proc. of the 2012 IEEE Int'l Conf. on Delovelopment and Learning and Epigenetic Robotics, 2012.

[13] M.E. Taylor, P. Stone, Transfer Learning for Reinforcement Learning Domains: A Survey, J. of Machine Learning Research, vol.10, no.53, pp.1633–1685, 2009.

[14] Y. Qiao, M. Suzuki, N. Minematsu, Affine Invariant Features and Their Application to Speech Recognition, Proc. of the 2009 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing, pp.4629–4632, 2009.

[15] Y. Choe, H.-F. Yang, D.C.-Y. Eng, Autonomous Learning of the Semantics of Internal Sensory States Based on Motor Exploration, Int'l J. of Humanoid Robotics, vol.4, pp.211–243, 2007.

[16] A. Censi, R.M. Murray, Uncertain Semantics, Representation Nuisances, and Necessary Invariance Properties of Bootstrapping agents, Proc. of the 2011 IEEE Int'l Conf. on Development and Learning, vol.2, pp.1–8, 2011.

[17] J. Tani, On the Interactions between Top-down Anticipation and Bottom-up Regression, Frontiers in Neurorobotics, vol.1, pp.1–10, 2007.