Intrinsically Motivated Anticipatory Learning Utilizing Transformation Invariance

Gakuto Masuyama, Atsushi Yamashita and Hajime Asama
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
Email: {masuyama, yamashita, asama}@robot.t.u-tokyo.ac.jp

***Abstract*—In this paper, novel reinforcement learning framework with intrinsic motivation to reproduce past successful experience is presented. Geometric transformation invariance is utilized to measure the reproducibility of experience. Top-down "expectation" to reproducibility of experience effectively biases strategy of exploration. As a result of consistent exploration via reproduction of successful experience, learning process is accelerated. Simulation experiments in grid world demonstrate useful characteristics of proposed framework.**

I. Introduction

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

A number of researches have been presented that show effectiveness of intrinsically motivational system implemented on machine learning algorithms. In [2], intrinsic reward for salient event to acquire useful skills is applied to option framework. Utilization of the temporally extended actions, i.e. option enables the agent to learn efficiently than a learner only with extrinsic reward. In [3], intrinsically motivational system called Intelligent Adaptive Curiosity (IAC) is presented to avoid stagnation of learning due to unpredictability in stochastic environment. IAC observes learning progress and try to maximize it. Thus the agent is pushed toward novel situation, and fully learned or unpredictable situations are avoided consequently. Such researches about intrinsically motivated learning methods demonstrated that appropriate utilization of task independent motivation could promote learning progress.

Previous studies have mainly focused on bottom-up process with respect to sensory input/feature vectors and their analyses to acquire useful skills or to speedup learning. However, it is known that bottom-up process and top-down process are used concurrently in information processing of human, which is one of a clear-cut objective of developmental robotics. Bottom-up process is data-driven: passive and sequential process based on continuous features obtained by input from peripheral system and analyses of them. On the other hand, top-down process is concept driven: active and explorative process guided by discrete signs representing abstract knowledge in cerebrum. Correspondence between the knowledge and result of analyses by bottom-up process is not necessarily clear. Top-down process generates "expectation" based on the knowledge, and it biases information processing system to realize expected results; it actively conceptualizes world model. The top-down bias on information processing system is associated with facilitation of object recognition [4]. Moreover, certain type of cognitive bias is related to motivation [5]. We take particular note of these unique aspects in top-down process related to intrinsic motivation, which is not very common in the context of intrinsically motivated reinforcement learning.

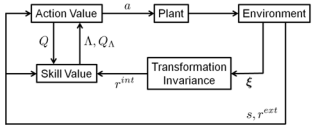

Fig. 1. Framework of proposed method

It is obviously important for design methodology of autonomous robot to consider bottom-up process such as skill acquisition in early stage of development. However, as the skill is accumulated and hierarchical structure is extended, unbounded extension of spatio-temporal search space should occur. As a result, apparent learning speed of agent would decrease as it develops. To this problem, implementation of top-down process on reinforcement learning is discussed in this paper; "expectation" based on knowledge and intrinsic motivation to the expectation offers efficient strategy of exploration. Simulation experiments demonstrate that biased strategy of exploration, which is exploiting the knowledge, could accelerate learning progress.

II. Methods

In Fig. 1, framework of proposed method is shown. Proposed method is constructed based on Q-learning [1]. Action value $Q(s, a)$ is learned by Q-learning, and skill value $Q_\Lambda(s, \Lambda)$ is learned in parallel. Skill $\Lambda = (A_\Lambda, M_\Lambda)$ is defined as successful experience in the past. Here $A_\Lambda = \{a_\Lambda(1), a_\Lambda(2), \cdots, a_\Lambda(T_\Lambda)\}$ is finite ordered action set, and $M_\Lambda \in \mathbb{R}$ is abstract representation of observed sensor information during execution of A_Λ. Thus skill value is learned in Semi-Markov Decision Process [6]. In action selection process (not action value update process), action value is instantaneously biased by skill value along skill execution.

Consequently, skill is adapted to a novel environment and embedded to action value. On the other hand, action value is utilized to update skill value. Thus the bias from skill value to action value works to reduce search space. If agent could move to better state as a result of skill execution, value of the skill would increase. Then the bias from skill value is also reinforced.

Additionally, intrinsic motivation to reproducibility of past successful experience is also introduced to skill value update law. The abstract knowledge of past successful experience M_Λ and abstract knowledge of present experience are relativized to measure the reproducibility. If the execution of skill demonstrates high reproducibility, the skill is motivated intrinsically. The abstract representation of experience is given by affine transformation invariant feature [7]. Affine transformation invariant feature could be generalized representation of observed data sequence. Suppose simple wall following behavior of robot equipping range sensor, it is not important to distinguish which side of the robot the wall exists. Geometrical structure in sensation space associated with wall following behavior is the only important thing in this case, because the two environments are observed symmetric for the translatory movement of the robot. Therefore the invariant property to geometric transformation could represent only essential information of temporal sensor readings if appropriate feature extraction is done. The intrinsic motivation is independent of specific task. However, it could provide consistent explorative strategy to the agent, because it motivates the agent to reproduce past successful experience. Consequently, bias from skill value operates as top-down process based on knowledge, and the internal model of the world (i.e. value functions) is conceptualized.

III. Results

Simulation experiments were conducted for simple navigation task in 2-dimensional grid world to emphasize fundamental effect of proposed framework. Simulated agent has four primitive actions; moving one cell to the up, right, down, and left directions. Environment is a square walled room (15×15). There are 45 obstacles that occupy one cell, but obstacles do not exist at cells bounding the wall. The agent starts at bottom-left corner and destination state is upper-right of the room. When the agent reach the destination state, it receives extrinsic reward 5; when the agent collides with obstacle, it receives extrinsic reward -1; and step penalty -0.1 for each primitive action is also imposed. Action and skill selection policy is soft-max action selection [1]. The agent is supposed to equip range sensor and odometer. In this paper, skill is acquired preliminary in smaller environment than tested ones. Q-learning is implemented and learned optimal action sequence and observed invariant feature are extracted as skill. With above settings, 50 trials were implemented. Every trial was tested with different obstacle and skill settings. For one trial, mean of 10 experiments was used for results.

In Fig. 2, learning curve of extrinsic reward is shown, and in Fig. 3, learning curve of intrinsic reward per one skill execution is shown. For each figure, horizontal axis represents

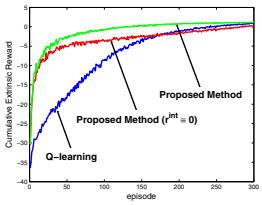

Fig. 2. Learning curve for extrinsic reward

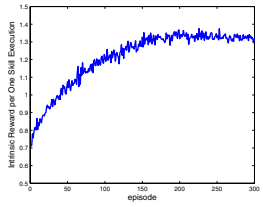

Fig. 3. Learning curve for intrinsic reward per one skill execution

episode. In Fig. 2, Q-learning, proposed method with no intrinsic reward, and proposed method is compared. There is prominent difference in learning progress, though update law of action value is the same among three methods. The skill offers consistent explorative strategy due to its nature as successful experience. Therefore proposed method and proposed method without intrinsic reward demonstrated better performance than Q-learning in early stage of learning. Additionally, there is difference between the two in convergence performance. In Fig. 3, acquired intrinsic reward increased in accordance with progress of learning. Skills possibly prevent learning unless the agent has internal criterion to evaluate the skill. Thus the intrinsic reward (or environment recognition via transformation invariance) plays a role of regulator against instantaneous bias to action selection process.

IV. Conclusion

In this paper, framework of reinforcement learning method based on intrinsic motivation to reproducibility of past successful experience is presented. Geometric transformation invariance is introduced to measure the reproducibility of sequential observations. The learner utilize the knowledge about past successful experience, called skill, to explore the search space with bias from the "expectation" of reproducibility. It offers consistent strategy of exploration, and intrinsic motivation promotes the biased strategy of exploration. Consequently, the top-down process via skill accelerates learning progress.

References

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

[2] S. Singh, A. G. Barto, N. Chentanez "Intrinsically Motivated Reinforcement Learning," Proceedings of Advances in Neural Information Processing Systems 17, pp. 1281-1288, 2005.

[3] P. -Y. Oudeyer, F. Kaplan, V. V. Hafner, "Intrinsic Motivation Systems for Autonomous Mental Development," IEEE Transactions on Evolutionary Computation, vol. 11, No. 2, pp. 265-286, 2007.

[4] C. Summerfield, T. Egner, "Expectation (and Attention) in Visual Cognition," Trends in Cognitive Sciences, vol. 13, No. 9, pp. 403-409, 2009.

[5] Z. Kunda, "The Case for Motivated Reasoning," Psychological Bulletin, vol. 103, No. 3, pp. 480-498, 1990.

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

[7] Y. Qiao, M. Suzuki, N. Minematsu, "Affine Invariant Features and Their Application to Speech Recognition," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4629 - 4632, 2009.