A Substitute Target Learning-based Inverted Pendulum Swing-Up Control System

SYAFIQ Fauzi Kamarulzaman, Takeshi SHIBUYA and Seiji YASUNOBU
University of Tsukuba
Tennodai 1-1-1, Tsukuba, 305-8573 Ibaraki, JAPAN
E-mail: syafiq_ics@edu.esys.tsukuba.ac.jp

***Abstract*—Human are able to perform a movement smoothly among constraints. This is because a human has the ability to place a series of substitute targets and produces a series of action based on those targets when the control purpose could not be achieve directly by a configurable control target. In this research, human ability to construct a series of substitute targets in order to achieve its control purpose is imitated and applied on an inverted pendulum swing-up control system. The pendulum uses substitute target knowledge to change the position of the cart simultaneously in order to swing the pendulum to achieve its final target which is an inverted state. The knowledge is constructed using reinforcement learning in order for the system to learn and select the most suitable target depending on state. The system effectiveness is later confirmed by simulations.**

I. Introduction

A conventional control method sets the control target beforehand based on a control purpose [4]. However, some control purpose is hard to achieve using a conventional method since a strong non-linearity exists between configurable control targets and the control purpose. Therefore, reinforcement learning is used as a method to produce an action that can yield the most reward upon achieving the control purpose [1] despite of configuring any control target.

Current research which uses reinforcement learning controls a control object using an output which was produced directly from the reinforcement learning table. This is difficult for the system to control through existing constraints. Therefore a control object needs to learn to produce a series of substitute targets to achieve its control purpose.

In this research, a control system using an alternative method of reinforcement learning is proposed by learning to produce a substitute target knowledge that helps the system to configure substitute targets to achieve its control purpose. An Inverted pendulum swing-up control system is constructed based on above method. The effectiveness of the system is later confirmed through simulations.

II. Control System Design

The proposed system shown in Fig. 1, consist of 3 major areas. These areas are the (i) Control Area, (ii) Recognition Area, and (iii) Knowledge Learning Area. The Control Area is used to change the output of substitute target displacement into voltage input for the control object motor. Recognition Area identifies the current cluster based on the current control object parameters. Knowledge Learning Area is the area where reinforcement learning takes place.

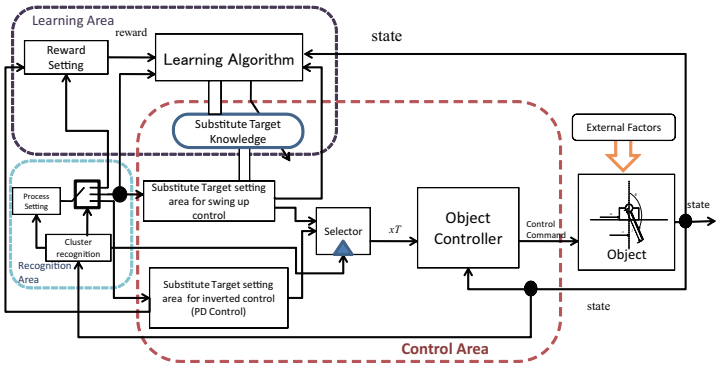

Fig. 1. The proposed system

A. Reinforcement Learning Algorithm

Reinforcement learning is proposed in order to implement a human-like knowledge into the control system. Therefore, Q-algorithm is used in order to construct a knowledge as shown in the algorithm below [1]:

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a')] \qquad (1)$$

where,
r: Reward for the state,
α: Learning rate,
γ: Discount rate.
In this research, Q-learning algorithm is used to construct substitute target knowledge.

B. Substitute Target Knowledge

The substitute target is constructed during a particular state several times to enable the system to propel the control object towards its final state. Thus, the distance between the control object current position and the substitute target is used in the reinforcement learning algorithm. The main reason for this change is for the system to consider any constraints around the control object, which in pendulum case is the cart position, x. Therefore, it is easier to learn using substitute target than to learn using controlled object voltage output.

The control object parameters will be identified as state, s which is use in equation (1). Instead of evaluating action, a,

the system evaluate the substitute target displacement, Δx in the value knowledge, Q which can be defined as

$$Q(s, \Delta x). \tag{2}$$

Thus, the substitute target, xT is generated by the sum of the control object current position, x_{now} and the substitute target displacement, Δx which can be written as

$$xT = x_{now} + \Delta x. \tag{3}$$

Therefore, the Q-algorithm can be rewritten to construct a substitute target knowledge as shown:

$$Q(s, \Delta x) = (1-\alpha)Q(s, \Delta x) + \alpha[r + \gamma \max_{\Delta x'} Q(s', \Delta x')] \tag{4}$$

The substitute target displacement, Δx is generated depending on the parameter state, s and selected among the index of Δx using $roulette$[4] and $greedy$[1] selection method.

III. Simulation

The simulation for the system is conducted on an inverted pendulum which is used as a control object as shown in Fig. 3 that is based on an experiment device in Fig. 2. The simulation is run for 2300 episodes where each episode consists of a 10 second run time. The specifications of the pendulum used in this simulation are as shown in TABLE I. Reinforcement learning occur during the pendulum in a full downward position, $\theta = \pi$ radian.

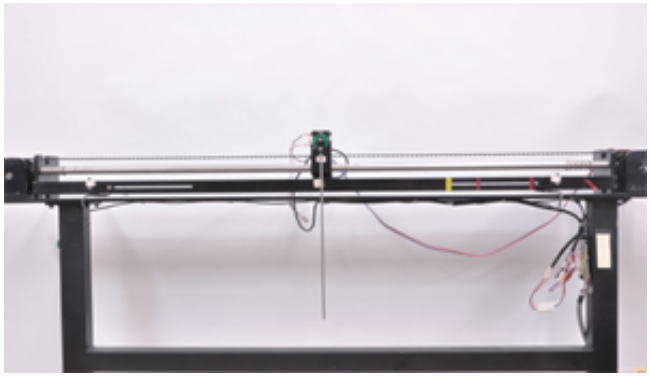

Fig. 2. The experiment device based for the simulation (Japan E.M. Co.,Ltd.)

The system will evaluate the most suitable substitute target based on the cart‘s movement. In this case, the substitute target is the sum of the carts movement and the substitute target displacement supplied from the knowledge table as mentioned in II.

The Q-learning parameters in TABLE II are used in II system’s learning algorithm. These parameters are previously selected.

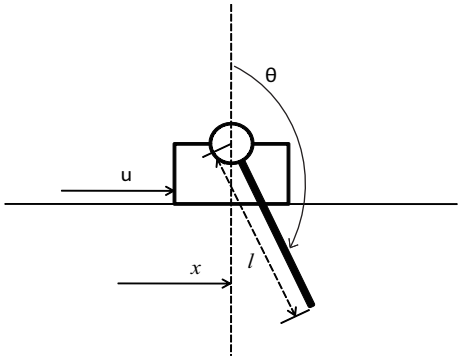

Fig. 3. Inverted pendulum diagram

TABLE I
SPECIFICATIONS OF THE INVERTED PENDULUM

cart mass	3.117 (kg)
Pendulum Length	0.4 (m)
Pendulum Mass	0.08 (kg)
Pendulum Inertia	$\begin{bmatrix} 0.0012 & 0 & 0 \\ 0 & 1.6 \times 10^{-7} & 0 \\ 0 & 0 & 0.0012 \end{bmatrix}$ (kgm^2)

TABLE II
Q-LEARNING PARAMETERS

Learning rate,α	0.5
Discount rate,γ	0.3

State Parameters	Range	Intervals
Cart Position, x	-1.0 $\sim$ 1.0 (m)	0.2(m)
Pendulum angular velocity, ω	-14 $\sim$ 14 (rad/s)	2 (rad/s)

action parameters	Range	Intervals
Substitute target displacement, Δx	-0.2 $\sim$ 0.2 (m)	0.05(m)

A. Control object movement state clustering

Movement state clustering is a method used to separate the pendulum state into several clusters. This method makes it easier to determine and view the process needed to control the object from one movement state to another. In the case of an inverted pendulum control, clusterization/clustering is made based on the pendulum angle, θ, and the pendulum angular velocity, ω.

As shown in Fig. 4, each cluster is given a name for easy recognition. These clusters are very important to determine the previous and the current cluster as the process can be selected according to the changes between these two clusters.

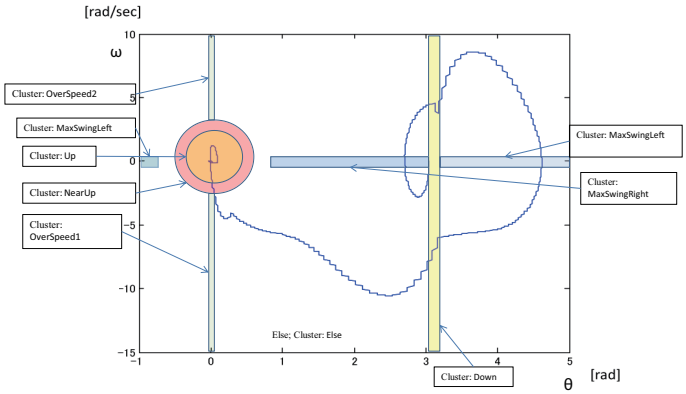

Fig. 4. Pendulum State Clusterization

B. Process selection via cluster change

In the Recognition Area, clusters will be classified according to Fig. 4. The previous cluster will be recorded within the system and it will be compared to the current cluster in order to select a suitable process. The process will be selected according to Fig. 5.

Previous cluster \ Current Cluster	Down	MaxSwingLeft	MaxSwingRight	OverSpeed1	OverSpeed2	NearUp	Up	Else
Down	21	x	x	x	x	x	x	21or 22
MaxSwingLeft	x	21	x	x	x	x	x	11
MaxSwingRight	x	x	21	x	x	x	x	11
OverSpeed1	x	x	x	21	x	x	x	12
OverSpeed2	x	x	x	x	21	x	x	12
NearUp	x	x	x	x	x	1	1	21
Up	x	x	x	x	x	1	1	1
Else	20	21	21	21	21	11	1	21 or 22

Process: 1 Stabilization Controller & Reward Setting: Reward = r
21 none
10 Reward Setting : Reward = 0
11 Reward Setting : Reward = r-Δr （Δr : discount rate)
12 Reward Setting : Reward =-r
20 Swing Up Controller & Table Update
21 none
22 Continuous from 20

Fig. 5. Process Selection

This method provides a clear view on how processes are determined according to the state cluster. The method also used to effectively configure necessary rewards for the reinforcement learning algorithm based on the pendulum state. Thus, it helps the system to effectively configure the series of substitute targets needed to control the object upon achieving its control objective.

IV. Simulation Result

In order to ensure a smooth simulation during the learning process, knowledge is previously constructed. This knowledge is based on an analysis of pendulum swing-up control. This is used to clearly detect any unexpected programming problems and to compare with results from a simulation using random knowledge. The simulation is done using $roulette$ selection for 2000 episodes before changing to $greedy$ selection for 300 episodes. $Roulette$ selection helps the system to include random selection of substitute target displacement unlike $greedy$ selection which only select a substitute target displacement with the highest value. This widens the range of learning inside the knowledge.

A. Comparison between simulation using constructed knowledge and random knowledge

The success rate for the system in both simulations can be seen in Fig. 6.

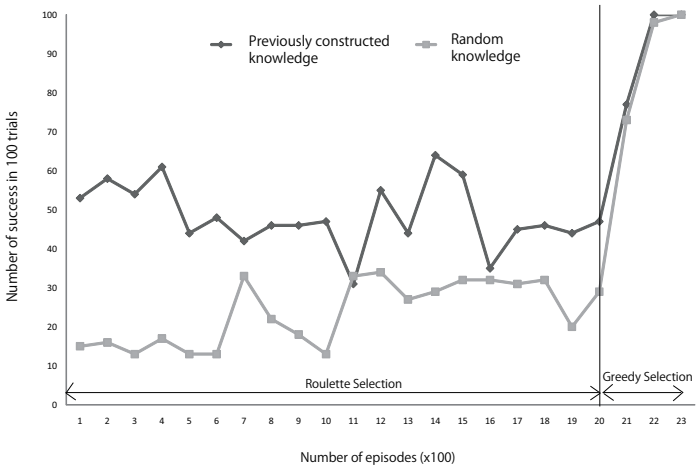

Fig. 6. Pendulum swing up control success rate

In the simulation which uses constructed knowledge, it can be seen that the success rate remains higher than the random knowledge through the simulation. Unlike the constructed knowledge, the simulation which uses random knowledge shows that the success rate increases through the simulation. This shows that the proposed system is able to learn to increase the success rate for achieving its control purpose. However, there is a gap in success rate between random knowledge and constructed knowledge. This is because the system may require more episodes than the amount done to achieve a similar success rate. After 2000 episodes, $greedy$ selection is used to boost the success rate to maximum.

B. Simulation using constructed knowledge as initial knowledge

Fig. 7 and Fig. 8 is the constructed knowledge during the initial state. At this moment, the constructed knowledge uses a previously analyzed knowledge which only based on pendulum swing-up control as mention before. Fig. 9 and Fig. 10 is the constructed knowledge result after the simulation. From these results, it is understood that there are changes/renewal within the knowledge. Therefore, it is assumed that the learning process runs according to expectation.

Fig. 11 shows the area of states which renewal occurs. Since the constructed knowledge is based on analyzed pendulum swing-up control, only few states were renewed. This is

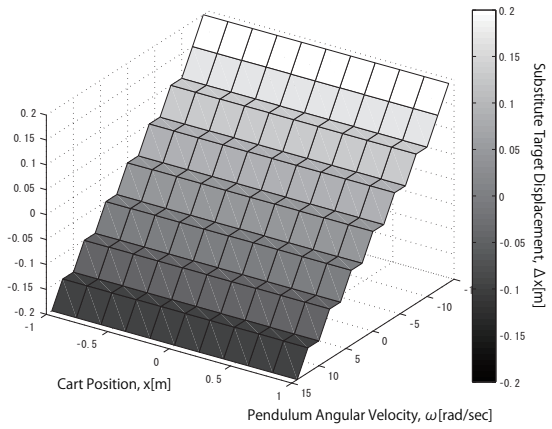

Fig. 7. Best substitute target displacement based on constructed knowledge state initial surface

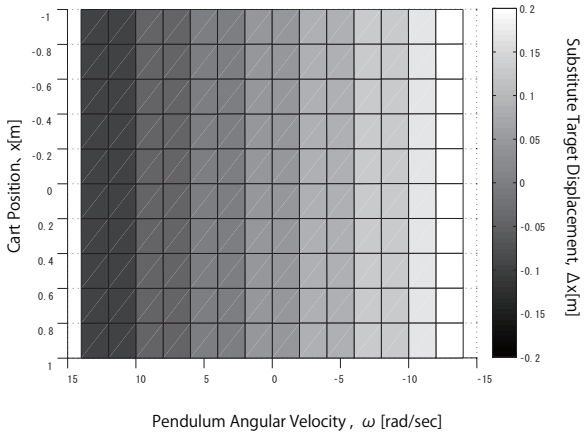

Fig. 8. Best substitute target displacement based on constructed knowledge state initial surface (above view)

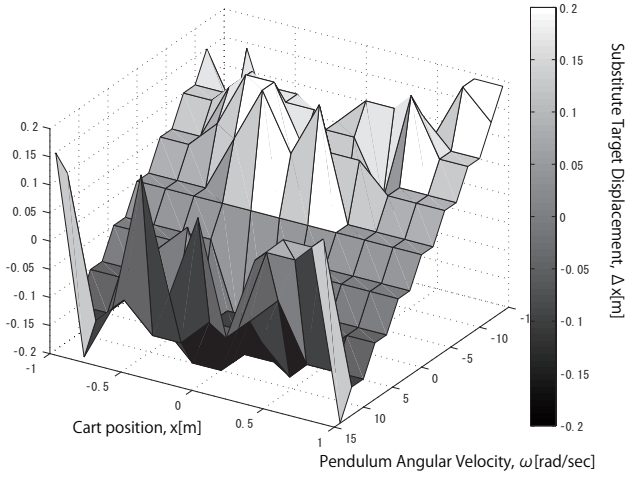

Fig. 9. Best substitute target displacement based on constructed knowledge state initial surface

because the system selects substitute target displacements which is known to be able to generate successful movements

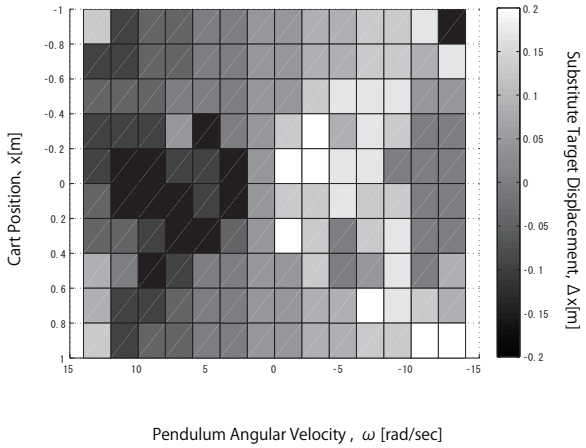

Fig. 10. Best substitute target displacement based on constructed knowledge state initial surface (above view)

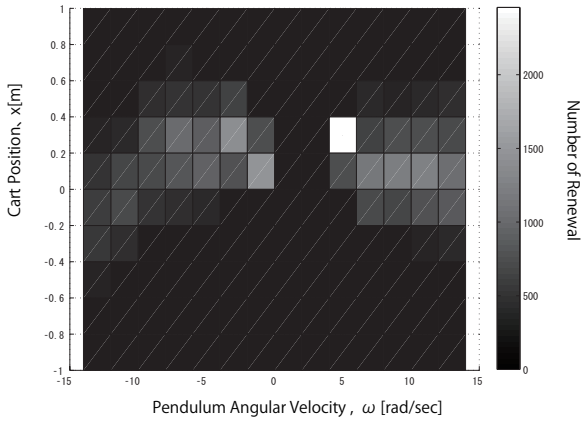

Fig. 11. Amount of renewal repetition according to state

by its value to achieve inverted state. Therefore, other state is less likely to be selected to avoid unnecessary movements.

C. Simulation using random knowledge as Initial Knowledge

Fig. 12 and Fig. 13 is constructed for a random knowledge during the initial state. From Fig. 12 and Fig. 13, random knowledge consists of random value for knowledge state against substitute target displacement. Fig. 14 and Fig. 15 is constructed from the random knowledge after the simulation.

Based on Fig. 14 and Fig. 15, the system learning process runs smoothly in order to achieve its target. However this knowledge is different from the first simulation. Wider range of states had been used during learning which can be seen from the range of state being renewed from Fig. 16. This is because by using random knowledge, random states has a random value which could make the system to select substitute target displacement from wider range of states.

In this case, wider range of states will construct a better knowledge which can even consider any constraints within their movement path. The constraints stated in cart position, x state parameters in TABLE II affect the knowledge as can be seen in Fig. 15. In Fig. 15, when cart position, $x = 1$

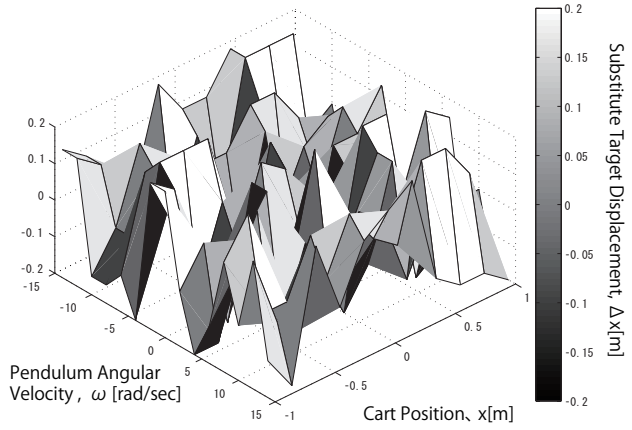

Fig. 12. Knowledge state and substitute target displacement evaluation initial surface

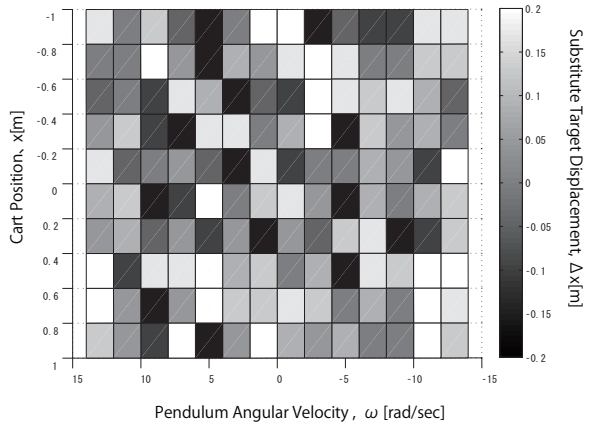

Fig. 13. Best substitute target displacement based on random knowledge state initial surface (above view)

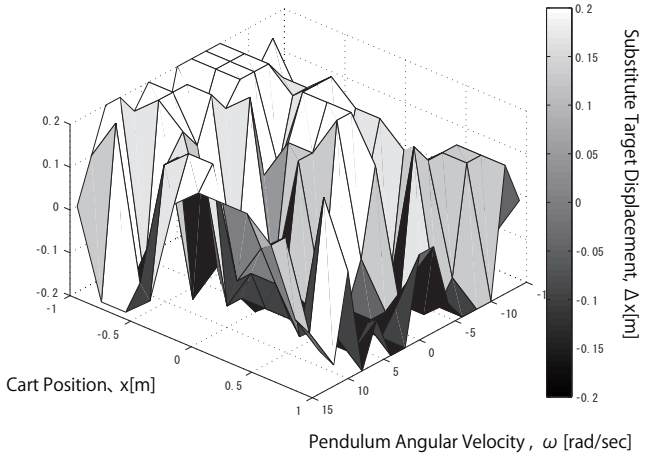

Fig. 14. Knowledge state and substitute target displacement evaluation after simulation

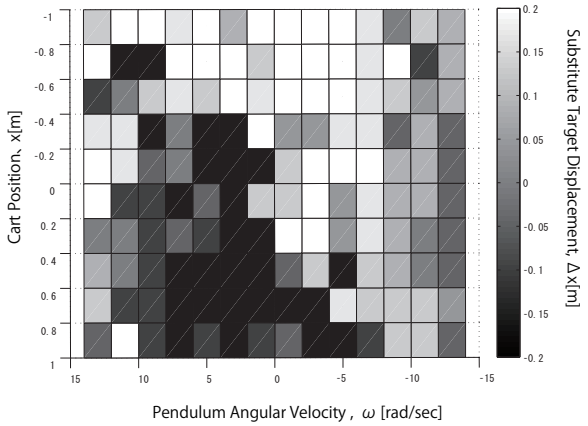

Fig. 15. Best substitute target displacement based on random knowledge state after simulation (above view)

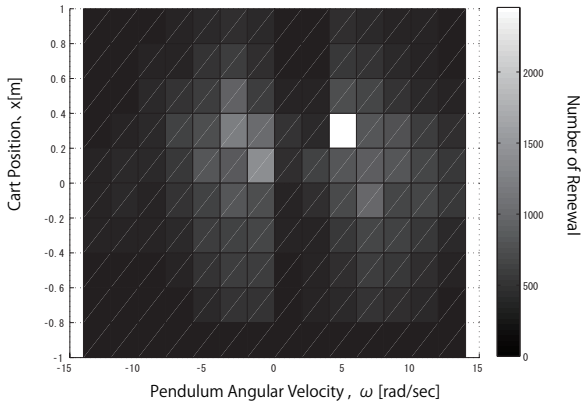

Fig. 16. Amount of renewal repetition according to state

m and $x = -1$ m, the system has the tendency to move the cart towards $x = 0$ m direction. Therefore, based on Fig. 15, it can be assumed that the system can consider any existing constraints within the cart movement path.

D. Successful swing up result

A successful swing-up result is taken at 2300th episode from both constructed knowledge and random knowledge simulation. At near 2300th episode, both simulations results repeatedly uses a constant method to swing-up the pendulum. However, both simulations results produce 2 different swing-up methods for the system.

1) Constructed knowledge successful swing up result: Based on Fig. 17, substitute target displacements are generated until up to nearly 3 second when the stabilization control occurred. From Fig. 17, it can be seen that there were 4 times when the pendulum angular velocity, ω increases and decrease at pendulum angle, $\theta = \pi$ radian. This shows that 4 substitute target had been generated which can be seen in Fig. 18.

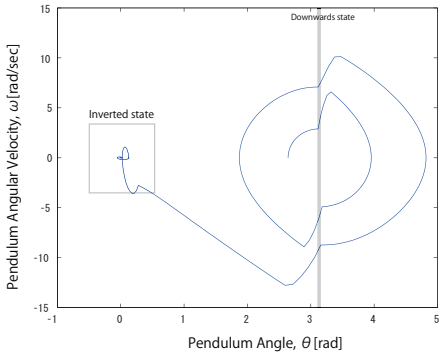

Fig. 17. Angular displacement and angular velocity relation from a successful swing up result (Constructed knowledge)

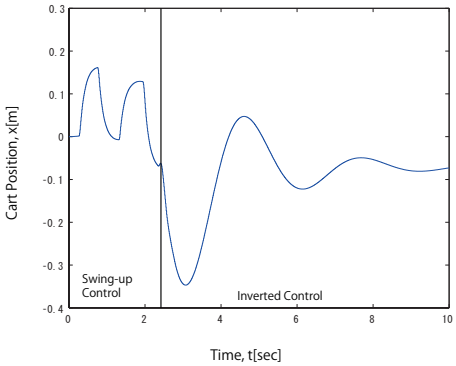

Fig. 18. The cart movement of a successful swing up result (Constructed knowledge)

2) Random knowledge successful swing up result: Based on Fig. 19, substitute target displacements were generated until up to later than 6 second when the stabilization control occurred. From Fig. 19, it can be seen that there were 7 times when the pendulum angular velocity, ω increases and decreases at pendulum angle, $\theta = \pi$ radian. This shows that 7 substitute target had been generated which can be seen in Fig. 20. However, there were 3 times when the system decided to maintain the cart position, x at pendulum angle, $\theta = \pi$ radian.

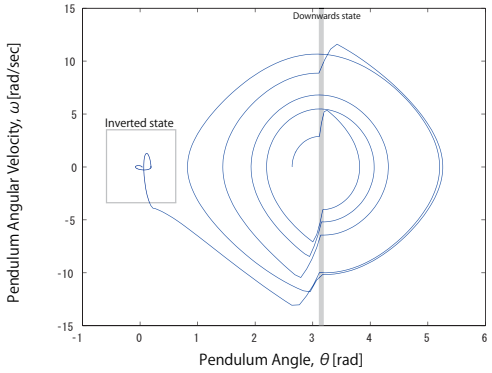

Fig. 19. Angular displacement and angular velocity relation from a successful swing up result (Random knowledge)

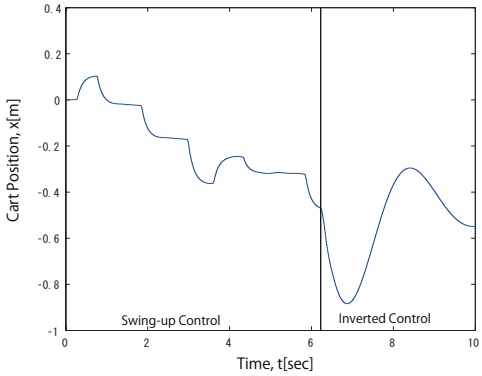

Fig. 20. The cart movement of a successful swing up result (Random knowledge)

V. Conclusion

In this research, a control system using an alternative method of reinforcement learning is proposed by learning to produce a substitute target knowledge that helps the system to configure substitute targets to achieve its control purpose. Reinforcement learning is used for the system to change or renew its own knowledge which will be used to generate substitute targets. It is understood that by using this method, the system could consider any constraints within its movement path by the result showed in the random knowledge simulation. The objective of this system which is to generate several substitute targets for the pendulum cart while swinging the pendulum towards its control purpose had been accomplished by using a constructed knowledge and random knowledge as its initial state. There are differences between these two results pendulum swing-up method produced at the end of these simulations. Although the success rate of the system during the random knowledge simulation is not as high as the constructed knowledge simulation in the beginning, it is certain that the system is capable to learn a suitable substitute target depending on the control object state even within existing constraints. Therefore, the proposed system using substitute target knowledge is capable of propelling control object using a series of substitute targets towards its control purpose.

References

[1] Richard S. Sutton, Andrew G. Barto,'Reinforcement Learning An Introduction' , MIT Press(1998).

[2] H. Iima and Y. Kuroe, 'Swarm Reinforcement Learning Algorithms Based on Actor-Critic Methods 2' , 35th SICE Symposium on Intelligent Systems , pp.9-14, 2008.

[3] E. Kawana and S. Yasunobu, ' An Intelligent Control System Using Object Model by Real-Time Learning ', SICE Annual Conference , pp.2792-2797, 2007.

[4] T. Matsubara and S. Yasunobu, 'An Intelligent Control Based on Fuzzy Target and Its Application to car like Vehicle ', SICE Annual Conference , 2004.