Design of Multiple-Functions Controller Based on Machine Learning

Graduate School of System and Information Engineering
University of Tsukuba

March 2015

Syafiq Fauzi Bin Kamarulzaman

Contents

List of Figures

List of Tables

Abstract

Human has the ability to learn and decide their action based on experiences when confronting a problem. Human decision often involves multi-functionality, where multiple control functions are applied for achieving a single goal. Conventional control often involves human in providing commands which mostly depends on the human decision. However, these decisions commonly involve single control function where multi-functionality can not be provided without human assistance.

Learning Control helps a machine constructs its own control knowledge autonomously through operation experiences. The development of Control Knowledge through Learning Control would require a period of training that could involve a number of failures among successful attempts. The Control Knowledge obtained is usually limited to single control function based on the training environment with less flexibility in varying environment.

Learning Control Systems with multiple functions could provide a wider range of control options against any environment. In this research, Learning Control System with multi-functionality is designed and developed. Here, application of Learning Control with multi-functionality provides a more human-like control operation with ability to adapt and consider the surrounding environment during control operation. The designs were evaluated through experiments and simulations where results confirm the effectiveness of the designed system. Through these results, the designs of multi-functions Learning Control may provide a safer and reliable control on control devices including complex non-linear control device.

Chapter 1

Introduction

1.1 Research Background

Human perform actions in order to complete task or react to surrounding environment. We render these actions in functions form. The actions are naturally based on purposes, which commonly act as goals. Successes and failures in achieving these goals are recorded in the human mind as knowledge, for references during future attempts. This form of learning represents human intelligence for being self-sustainable that is important in improving our skills for solving surrounding problems.

Applying such intelligence in machines has been an issue surrounding many researchers. Methodologies for self-sustained autonomous machines have been well developed and various new methods and ideas are continuously being proposed in order to reduce human intervention in managing these machines. Providing actions of machines in form of functions help machines to self-evaluate their actions. Human-like functions are one of the focuses of these methods and application may provide methods for self-sustained autonomous machines that could react and adapt to surrounding environment.

1.1.1 Multi-Functionality

Human functions are not limited to individual components where each functions only reacts to a single goal. A goal may require multiple human functions to be obtainable. For example, in case of hurdle race, two human functions of *jumping* and *running* are combined to cross the finishing line which acts as a goal. Here, multiple functions are utilized, where a professional with only either *jumping* or *running* skills are not certain to be capable of achieving the finishing line perfectly. The above ability here is described as $Multi-$

functionality. Through Multi-Functionality, an action can be learned and decided by multiple knowledge of skill, and applied when confronting a problem that cannot be solving by a single function. Here, Multi-functionality can be described as a quality of utilizing multiple functions for performing a single goal.

A device with multi-functionality could render an action that considers multiple characteristics in surrounding environment through application of knowledge of skills from various environments. A device with conventional control method only utilizes control command that produces action based on a single function. Method of self-sustained machines could only utilize a single function to become sustainable and lack of flexibility in confronting foreign characteristics simultaneously. Multiple control option is needed in self-sustained machines in order to become autonomous. Multi-Functionality may provide a wide range of control option against any environment in self-sustainable machines.

1.1.1.1 Multi-Functionality against Non-Linearity

Most control method considers linearity in a device for deciding control option. A device with non-linearity will not able to utilize a single control method for the entire system due to parameters that would render the system unstable at a certain state. For example, a pendulum-cart device has two different states that require different control methods for operation. Multiple functions are needed to manage these multiple states. Conventional control method such as Cascade PD Control can only provide two functions for swing and stabilization control. In case of more functions required, such method could not manage to perform successfully.

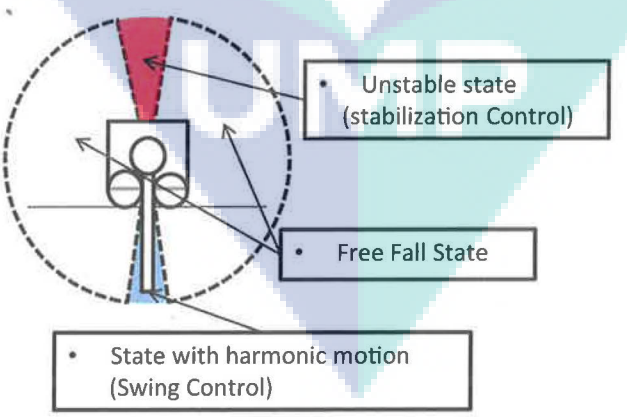


Figure 1.1: States for control of Cart Pendulum System.

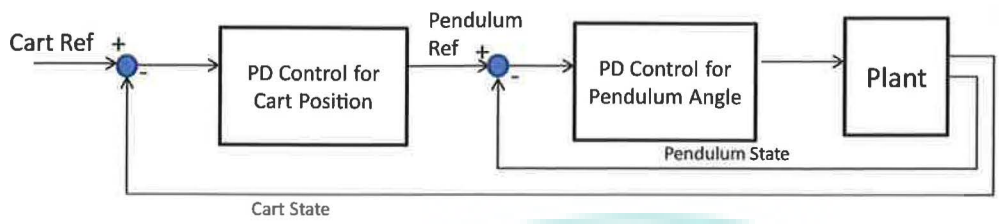


Figure 1.2: Cascade PD control of a Cart Pendulum System.

Non-linearity also exists in our common devices such as vehicles. Non-linear Control in machines is complex and hard without an expert human knowledge in the control system. Aerial hovering vehicles such as helicopters require multiple functions for managing multiple states using the Thrust and Cyclic.

Figure 1.3: Example of aerial hovering vehicle with non-linearity. (Parrot inc.)

Manipulation of angular orientation with thrust can provide position transition but requires skills in multi-functionality. Human multi-functionality provides expert control of machines with non-linearity. Providing multi-functionality in a non-linear control system could provide a safe and reliable control as good as an expert human.

Human multi-functionality provides expert control of machines with non-linearity due to utilization of multiple knowledge of skill when managing the machines. Through skills of angular orientation and hovering thrust manipulation, expert human pilots are able to perform radical movement of such machines in precision, for example, during position transition of the vehicles. They may react to surrounding environment while still maintaining stability of the machine that is easily affected by unstable states. Therefore, providing quality of multi-functionality as well as human-like functions in a non-linear automatic control system could provide a safe and reliable control replacing an expert human.

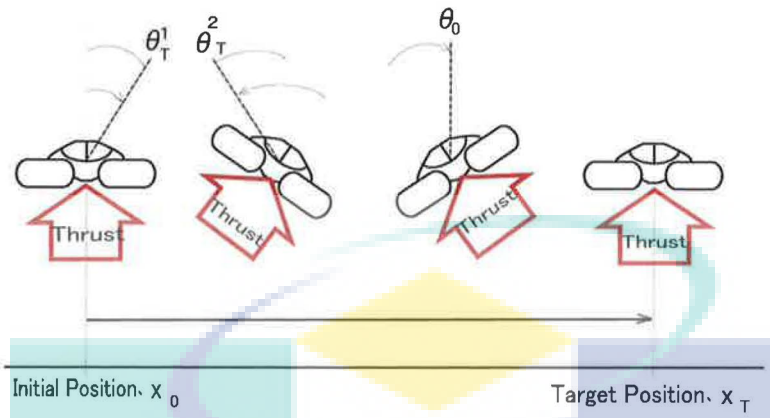


Figure 1.4: Control operation during position transition of aerial hovering vehicles.

1.1.1.2 Embedding Human Knowledge for Multi-Functionality

Embedding human like functions in a system through application of Intelligent Control that provides detailed decision during control operation in a certain environment. Various method concerning intelligent control system may help provides control alternative to an expert skills in controlling a device. Intelligent Control System provides autonomous development of control knowledge together with autonomous development of control strategy on a device. Control Knowledge and Control Strategy are developed depending on a human control decision together with the environment feedback. The developed Control Knowledge and Control Strategy may perform as well as an expert human controlling the machine, reducing the command burden on the human. However, the embedded human functions in the control knowledge are usually constrained to a single function.

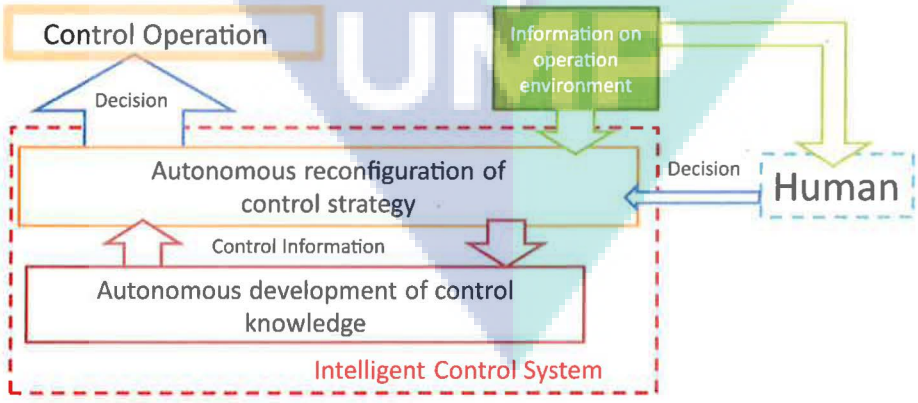


Figure 1.5: Structure of functions in an Intelligent Control System.

Learning is one of the qualities for developing control knowledge in Intelligent Control System. Control knowledge may be developed through learning method such as trial and error processes. Machine Learning provides option in generating development of control knowledge in an intelligent control system. Control knowledge may be developing through experiences by method in Machine Learning such as Reinforcement Learning. Development of control knowledge helps an Intelligent Control System remain self-sustained and adaptable to changes in surrounding environment. Therefore, new functions may be learned through the learning process giving quality of multi-functionality to the Intelligent Control System.

1.1.2 Learning Control

Learning is generally defined as the process of acquiring new knowledge. The process of acquiring new knowledge needs one to represent the knowledge in some form, as learning is constructing or modifying representations of what is being experienced [4]. The representations meaning varies depending on the knowledge it represents which can be in a form of algorithm, simulation models, control procedures and such.

The term of Machine Learning is derived by the ability of a machine on acquiring knowledge from experiences or a set of data. Mitchell [1] defines learning as performance improvements at some tasks through experience. Mitchell defines it precisely as,

> *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

To have a well-defined learning problem, three features concerning class of tasks, the measure of performance to be improved, and the source of experience must be defined. Thus, Machine Learning aims to have a computational mechanism that can learn to improve knowledge through operational experience.

1.1.2.1 Reinforcement Learning

Reinforcement Learning is known as trial and error style learning process that learns to map situations and actions by maximizing a numerical reward signals [4]. All Reinforcement Learning agents may have explicit goals. Using its experience, the agents improve its performance over time. Aspect of their environments can be sense and actions are

changeable to influence their environment. Reinforcement Learning acquires action rules for adapting with the surrounding environment. Reinforcement Learning operates through interactions and acquires knowledge by categorizing actions using rewards, optimizing the best possible action required in order to complete a task [3].

Reinforcement Learning normally consist four main sub-elements in its system [4]. A *Policy* to determine behaviour, a *Reward Function* to determine reward, a *Value Function* to emulate knowledge and sometimes a *Model Environment* to mimics the property of the environment. The relation between these elements can be seen in Figure 1.6.

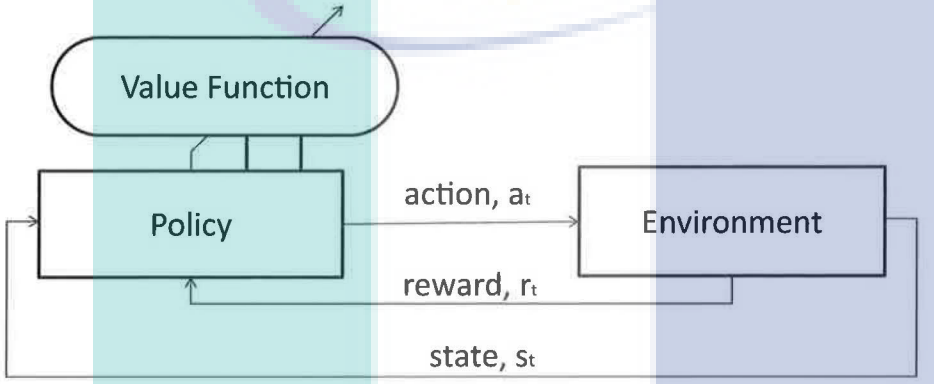


Figure 1.6: Interaction between policy, reward function and value function.

A Policy defines the agent behaviour. Policy perceives state mapping of the agent environment to actions to be taken when is those states. A Policy might be a simple function or a lookup table but sometimes involves extensive computation such as search process. Policy is the core component of a Reinforcement Learning agent since it alone determines the behaviour of the agent. A Reward Function defines the goal for the agent. Reward Function maps each perceived state to a single number which known as reward, indicating the desirability of the state. The purpose of Reinforcement Learning is to maximize these rewards in an operation. In other words, Reward Function defines the good and bad of an action for the system to operate. Reward Function is needed to alter the policy. Generally, actions with low reward will less likely to be selected by the policy repeatedly.

While Reward Function indicates good and bad action immediately, a Value Function acts as knowledge of the good and bad action experienced in a long term operation. The Value Function represents the value of states and indicates the desirability of the state reoccurrences in a long term operation. A state may have low rewards but high in value since it is regularly followed by other state that can yield high rewards. Therefore, a

method of converging the value of state-action pairs $Q(s_t, a_t)$ into an average is called the *Q-Learning* algorithm as

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)). \tag{1.1}$$

Reinforcement Learning may provide the ability of learning for a control system, for example, on a mobile robot. Programming all possible tasks for the robot can be hard and difficult, simplified by applying a learning ability for the mobile robot. In some cases, a control system would have problems with wear and tear in the control object hardware that can cause unprecedented misconfiguration during long term operation. Provided an ability to learn, the control system may adapt to the condition of its control object on any uncertainty and unforeseen changes by continuous self-calibration.

Learning control refers to the process on developing control strategy in a particular control system by trial and error [6]. This is a branch of Reinforcement Learning in control application where agent learns by analysing good and bad influences those results from its own action during control operation. Learning control resembles the way that humans and animals learn to construct their knowledge of movement strategy based on interaction with the environment.

1.1.2.2 Absence of Multi-Functionality in Reinforcement Learning

Through Learning Control, control knowledge of a control function can be created through the training by Reinforcement Learning. However, conventional Reinforcement Learning method does not provide application of more than one control function within a Learning Control System. Execution of more control function within a Learning Control System would require application of multiple learning processes within a control system. Methods concerning application of learning processes in Learning Control vary depending on application of the control device and the purpose of the system.

Multi-agent Reinforcement Learning is one of the method concerning application of multiple learning process within a Learning Control System. Application of multiple agents in Reinforcement Learning utilizes learning process for multiple agents, where these agents interact between each other in developing the desired control knowledge [49]. State transitions in the case of multi-agent Reinforcement Learning are the results of the joint action that was performed by the agents within the system. Rewards are evaluated through the joint action, and the control knowledge is updated through a joint policy. In this case, the goal

can be determined through adaptation of the dynamic behaviour between these agents [27]. In case of controls, through multi-agent Reinforcement Learning, dynamic behaviour of the agents performs an action that requires the agents to adapt through an environment but the functionality of these agents is limited [28]. Such behaviour, may have exploration task, where the agents has a function of maintaining a group of moving targets within the sensor range [50] [51]. In overall, multi-agent Reinforcement Learning only focuses on application of multiple agents by Reinforcement Learning for utilization of a primary function.

Hierarchical Reinforcement Learning applies a learning process for improving the reliability of Reinforcement Learning application in real-world problem. Conventional Reinforcement learning methods provides solution in providing adaptable control knowledge in form of value functions. However, the bigger the size of state-space variables, the performance of Reinforcement Learning reduces and would require a large scale of computational effort for the update of the control knowledge. Hierarchical Reinforcement Learning accelerates the reliability of the learning process, where state variables are independent from one another, ignoring irrelevant aspects when solving a sub-task [42]. Hierarchical Reinforcement Learning provides a form of decision management in a system, where sub-task will be surveyed by parent task, providing only relevant action depending on the sub-task performance. In this case, value function of the parent task is separated into value functions of sub-task, where learning process occur ignoring irrelevant sub-task during a precise operation. The value functions of sub-task are then converged into performing a value function of parent task [43]. The main purpose of such method is mainly to increase reliability of the learning process in a certain function that requires monitoring of multiple states in a more accelerated pace.

Applying Learning Control System with a number of Control Functions could provide a wider range of control option against any environment. The Control System should be able to develop and apply the required Control Function according to necessity and could provide a more versatile control operation. Current Reinforcement Learning does not emphasize multi-functionality in a control system. Therefore, a method of applying a number of control knowledge with decision management that can provide cooperation between each provided control function is deemed necessary for a quality of multi-functionality in control system.

1.2 Research Objective & Contents

Through multi-functionality and Learning Control, an idea of a control system that is capable of utilizing any functions while being self-sustainable is possible. Researches concerning Multi-functionality in Learning Control are unfolded in this dissertation. This dissertation provides control design that makes use Learning Control in providing multi-functionality in a Control System.

1.2.1 Research Objective

The objective of this research is to design and develop methods of applying Learning Control that provides multiple control function in control command autonomously during a control operation. Through this research, a control system that is self-sustainable, reliable and adaptable to its surrounding environment motivates the development of methods in achieving Multi-functional Learning Control System. Characteristics of such system can be divided into three qualities.

Firstly, the system is believed to be able to provide safe and reliable control operation in any environment through development of the control knowledge according to successes and failure during control attempts. Experience from past control attempts can be referred to while safer future attempts are being planned. Consecutive attempt continues the development of the control knowledge that renders the system upon becoming an expert system with expert control knowledge.

Secondly, the system is believe to be able to reduce dependency on human intervention by self-sustaining system development during control operation in a certain environment. Control Decision can mostly be provided by the system based on the control knowledge developed, reducing the need of human commands. Thus, reduces the requirement on skills on the human operators while maintains the expertise in executing the control operation.

Thirdly, the system is believed to be able to provide decision management in a Learning Control System, that could considers multiple functions during execution. Here, the system may provide wide range of control options during control operation while considering changes in surrounding environment.

The above characteristics provide ideas in designing systems that reflects the motivation of this research. Design of systems that consists above characteristics is unfolded in this dissertation in three chapters.

1.2.2 Research Content

Here, three phases of development were organized to fulfil the objective of applying multi-functionality in a control system. First, for applying multi-functionality in a non-linear system, a Substitute Target based Learning Control System with Multiple Control Function was design. Secondly, for applying human like multi-functionality in a control system, Learning Control System with multiple control function by multiple source of control knowledge was designed. Finally, for applying human like decision management with multi-functionality, Learning Control System with multiple control function by Compound Function was designed.

In this chapter, the background of the research is explained, concerning motivation in application of multi-functionality in controls by Learning Control System. Later, background research concerning Learning Control System is introduced, which emphasizes lack of focuses in application of Learning Control concerning multi-functionality. This leads to the objective of this research which explains the needs and potential of a Learning Control System that emphasizes on multi-functionality.

In chapter 2, a design of Learning Control System with multiple control function that applies substitute target for multi-functionality is introduced and applied on and cart-pendulum control system. The designed System focuses on providing multi-functionality in the pendulum swing up control that may considers surrounding constraints for achieving the inverted states. The system applies Learning Control in producing substitute targets for the cart position transition which swings the pendulum simultaneously. The substitute targets act as intermediate targets that help the system considers optimal cart movements to provide swinging motion on the pendulum that propels it towards the inverted states under the influence of environmental constraints.

In chapter 3, a design of Learning Control System with multiple control functions by multiple source of control knowledge is introduced and applies on control systems of cart-pendulum and aerial hovering vehicle. The design focuses in applying multi-functionality through application of multiple source of control knowledge. It was applied on rapid position controls of aerial hovering vehicle that was simulated through cart-pendulum controls. The design was improved for control of aerial hovering vehicle among constraints that was applied on simulation of aerial hovering vehicle. The designed utilizes multiple sources of control knowledge for providing controls of angular orientations on the aerial hovering vehicle.

In chapter 4, a design of Learning Control System with Multiple Control function by

Compound Function is introduced and applies on control system of a mobile robot. The design focuses in applying multi-functionality through application of multiple source of control knowledge that merges through utilization of Compound Function. It was applied for position transition and obstacle avoidance control of the mobile robot that was simulated and later applied on a real world operation. The design utilizes Compound Function for creation of Compound Knowledge that consists of compounded control information from the sources.

Finally, the designs in this research are concluded together with suggestion of further research.

Chapter 2

Multiple-Functions Learning Control by Substitute Target

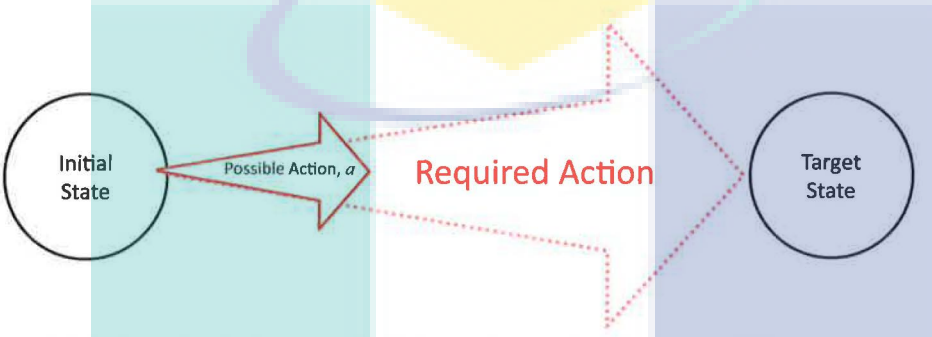
Designing Learning Control with a quality of multi-functionality requires recognition of continuing states during controls operation. Like human recognizing positions for the next step during walking, the positions of those steps reacts as substitute targets where the main target is the desired location of the human. The substitute targets provide options of manoeuvre, where certain action, in case of walking, can be operated flexibly along constraints during the manoeuver. Therefore, one of the designs concerning Learning Control with multiple functions involves application of substitute target in the Learning Control System.

2.1 Substitute Target

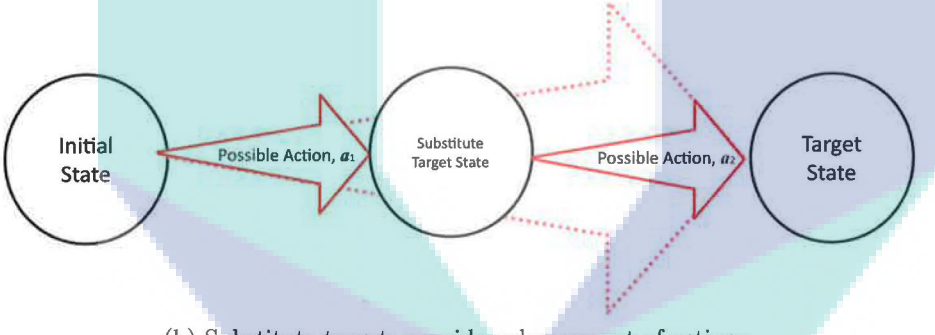
Conventional Reinforcement Learning involves application of state-action pair for providing control knowledge of a certain control operation. Optimum action is learned based on the states of the control object through the success and failure attempted during the control operation. Comparing such application to human, human decide a target or goal before applying an action. For example, in case of walking, a target for steps is determined before the action of walking is applied. A wrong position would render the walking operation colliding with constraints, or heading in the wrong direction. Targets make configuration easier, since target is a part of state elements, such as steps to location of human. Most controls of actuators apply targets as reference for feedback during control operation as well. Multiple target states provide multiple choices of actions for achieving goal and such supporting target states is defined here by substitute targets.

Substitute target is necessary for flexibility in providing system respond to the change of

situation in environment. Controls by substitute target provide flexible action in which important for having an adaptable Learning Control System for such application on machines with non-linearity. In this case, substitute target provides enhancements of action through continuity in applying those targets. For example as shown in figure 2.1, generating an initial action a_1 and continues with action a_2 during a control operation provides continuous action, or an action with increasing magnitude. Such function may provide precision of a higher magnitude action and reduces the risk of rampaging actions.



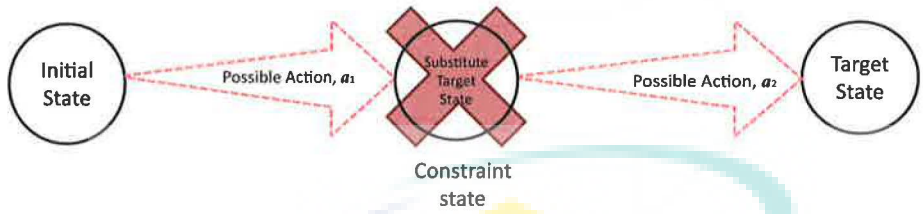
(a) Action with higher magnitude is required under limited possible action.



(b) Substitute targets provide enhancement of actions.

Figure 2.1: Substitute target provides continuity of action by providing an intermediate state.

Substitute targets may also provides rearrangements of control manoeuvre for adapting with constrained environment. During operation in a constrained environment, interference by constraint state would jeopardize the control operation, where rearrangement of controls manoeuvre are necessary. Figure 2.2 referred to a case, where substitute target provides rearrangement of actions, creating more substitute targets that provides a safer manoeuver for the control device. When one of the substitute targets are in a constraint state, a new substitute targets can be arrange to provide alternative for the required action. The arrangement of those substitute targets may vary depending on possible combinations that



(a) Constraint states interfere with continuity of action.

(b) Rearrangement of substitutes target helps avoid the constraint states.

Figure 2.2: Substitute target can be rearranged to satisfy the need for successful control manoeuvre along constraint states.

provide the required action for fulfilling the goal.

Utilizing substitute target provides flexibility in producing actions in control operations through Learning Control. Safer and more reliable control operation is possible through the application of substitute target in a Learning Control System.

2.2 Utilization of Substitute Target

Utilization of substitute targets may provides a safe and reliable control option for machines with non-linearity. Here, a control system that utilizes substitute target was designed to provide multi-functionality on machines with non-linearity for safe and reliable control operation. Substitute target was applied on a Learning Control System for cart-pendulum device, shown in figure 2.3. Application of such device requires three basic system functions in the designed system; the control function, learning function and recognition function.

In the Learning Control System designed, control function configures the control output for applying forces to the cart based on the targets instructed either from a policy of reinforcement learning or PD control. Learning function updates the knowledge of substitute

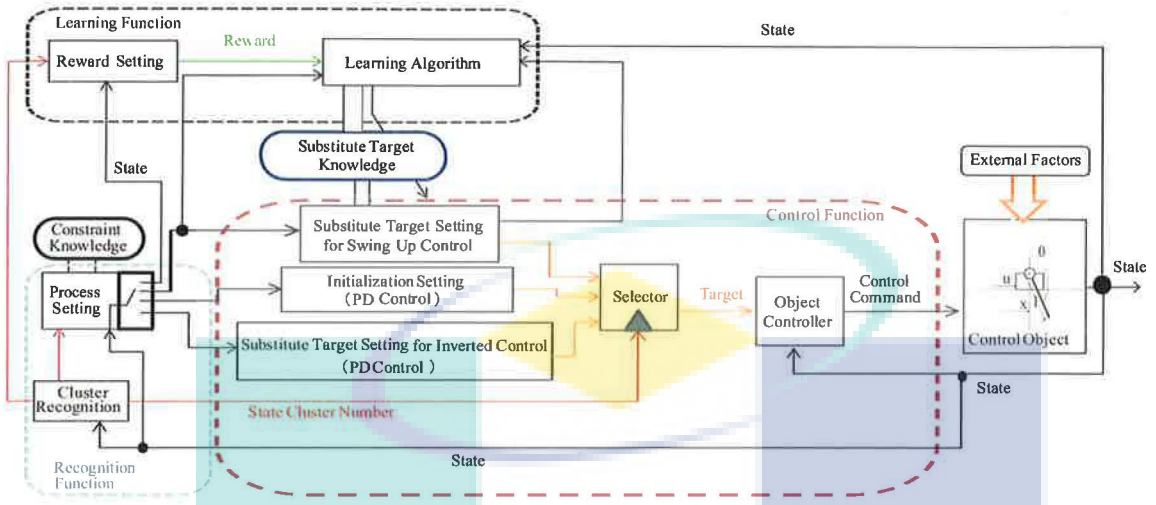


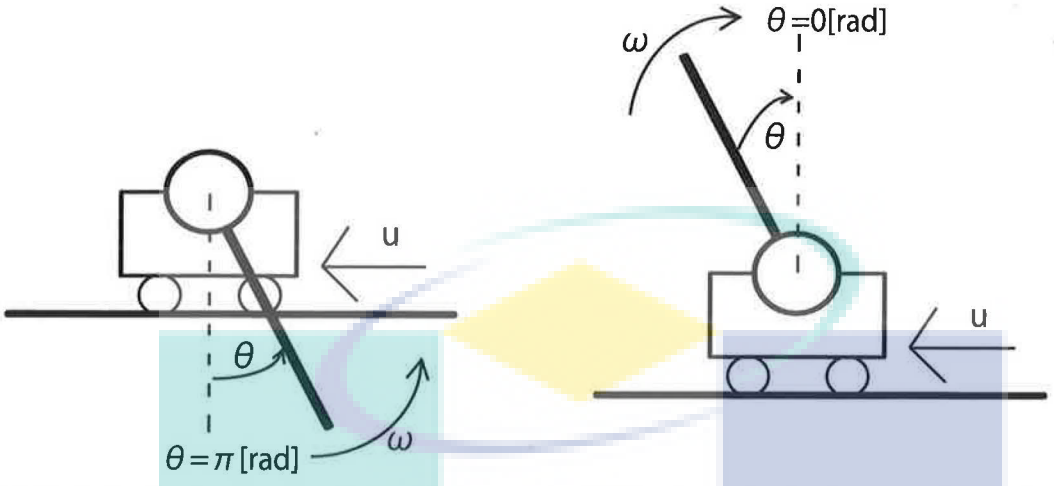
Figure 2.3: System structure for substitute target application.

target based on the reaction of the pendulum when applying the force to the cart. The recognition function determines the necessary control action depending on the states of the pendulum and the cart, in order to instruct the next required process. Interaction between each functions provide application of substitute target in controlling the cart for applying swing motion on the pendulum towards the desired goal state.

2.2.1 Control Function for Substitute Target System

Control function provides control options for the system to apply on the cart. The controls within the function consists two methods; Swing Control and Stabilization Control. Swing control generates forces for increasing the pendulum swing angle when the pendulum is in downward state. The stabilization control generates forces for decreasing the pendulum swing angle when the pendulum is near to inverted state.

During swing control, control output u provide forces to move the cart for increasing the swing angle of the pendulum. The cart moves to either right or left based on the pendulum angle θ and pendulum angular velocity ω for intensifying the pendulum swing, increasing the pendulum angle θ. The initial state of the pendulum was assigned on the downward position where the pendulum angle $\theta = \pi[rad]$ as shown in figure 2.4a. The pendulum angle θ will increase as the cart moves consecutively until approaching the inverted state. The Learning Control for substitute target was applied on the pendulum swing control. The swing up control arranges targets for cart movement and apply force u according to those targets.



(a) Control output u increases the pendulum angle θ during swing control.

(b) Control output u decreases the pendulum angle θ during inverted state.

Figure 2.4: Swing and stabilization control of the cart-pendulum system.

The stabilization control occurs when the pendulum approaches the inverted state. During stabilization control, the pendulum swing will be attenuated towards the inverted state as shown in Figure 2.4b. The cart will move to either left of right reducing the pendulum angle θ to $\theta = \pi[rad]$. Here, the occurrence of pendulum stabilization control and inverted state will be the goal for the learning control. The stabilization control was conducted and designed based on PD control.

Applying Learning Control by substitute target into a pendulum control system requires three major sections for controlling the cart movement in the Control Function. These sections are (i) swing up control section, (ii) inverted control section and (iii) initialization control section, that provides control command u for the cart.

The swing up control section provides control command for the pendulum during the pendulum downwards position. The control command is based on targets on the cart position axis, x. Substitute target displacement Δx is selected from the substitute target knowledge, $Q(s, \Delta x)$ which defined by value function Q for substitute target displacement Δx based on state s. Substitute target x_T was arranged during the pendulum downwards position based on the substitute target displacement Δx provided by the substitute target knowledge, $Q(s, \Delta x)$. Substitute target x_T was arranged based on the selected substitute target displacement Δx to the current cart position x_{now} as

$$x_T = x_{now} + \Delta x. \tag{2.1}$$

The inverted control section applies control using substitute target x_T arranged as in equation 2.1. This control section provides control when the pendulum approaches the inverted state as shown in Figure 2.4b based on the movement shown in Figure 2.5. However, substitute target displacement Δx was arranged through PD control after the pendulum reached the inverted state.

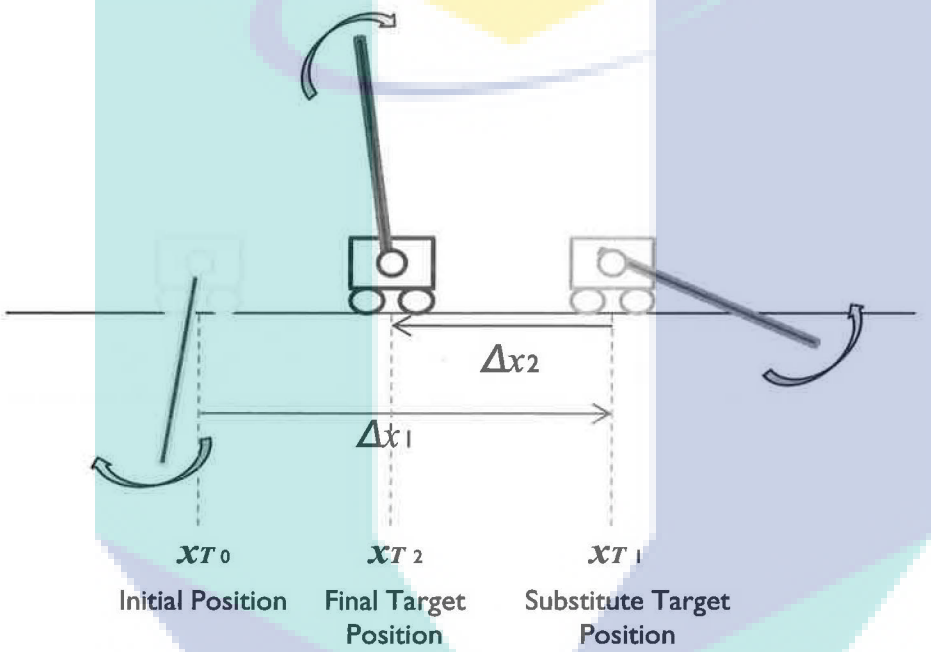


Figure 2.5: The swing control of the pendulum based on substitute target.

The initialization section provides control commands for moving the cart towards the initial position. This occurs after the pendulum achieved the inverted state or after any constraints encounter. This section controls using substitute target x_T, arranged based on a substitute target displacement Δx generated through PD control similar to the inverted control section.

2.2.2 Recognition Function for Substitute Target System

The recognition function assigns states of the control device into sets of state clusters. Certain range in state parameters is divided and separated into clusters for easy recognition. Here, pendulum cart position x, pendulum angle θ, and pendulum angular velocity ω were

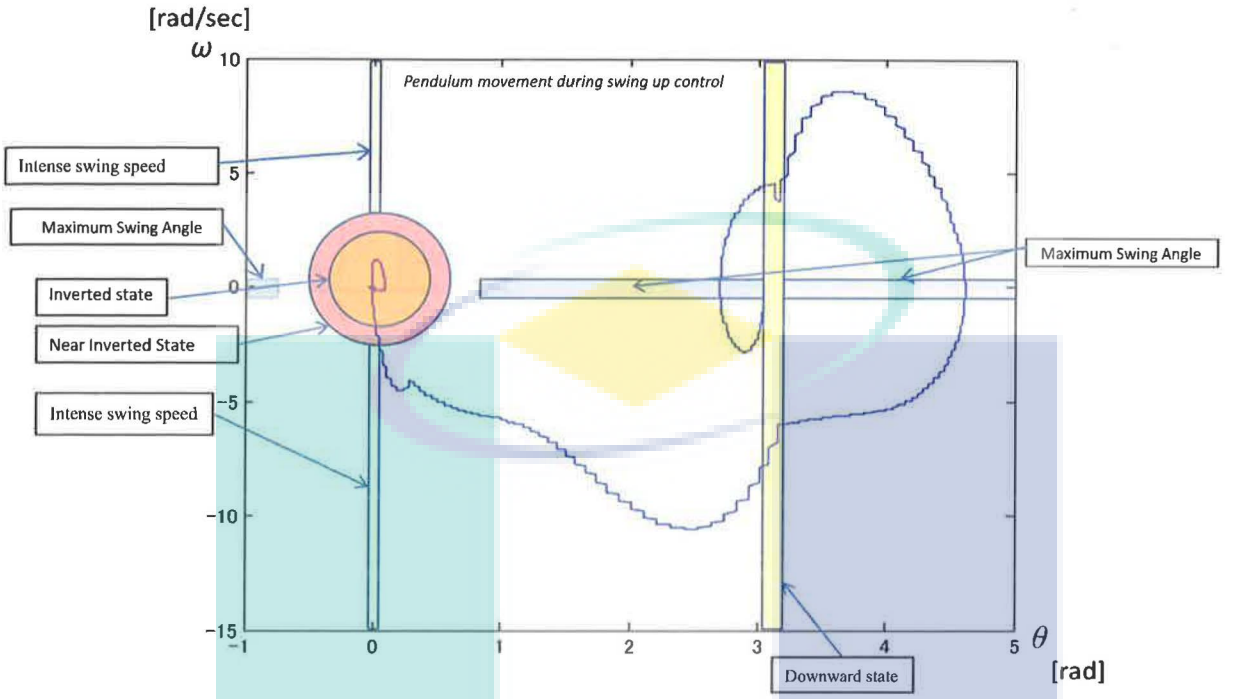
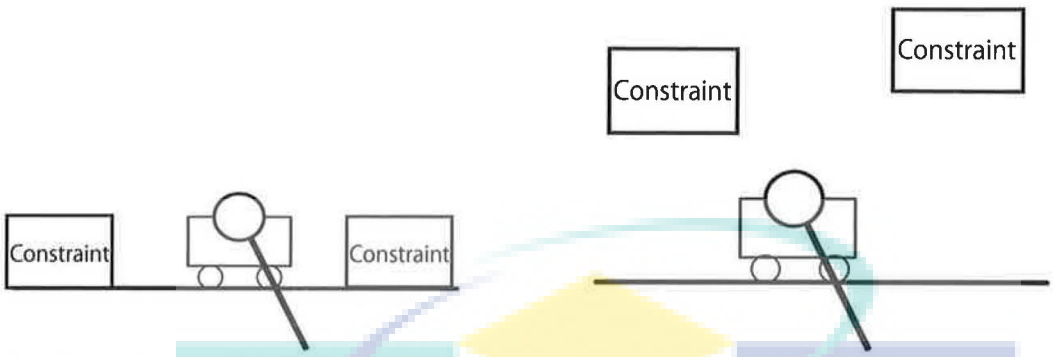


Figure 2.6: State clusters created from pendulum angle and pendulum angular velocity.

assigned in cluster of states as shown in figure 2.6. The recognition function applies duties to other functions based on these state clusters. It provides commands for control function for selecting suitable control sections for generating outputs. The state cluster provides determination of rewards based on the current state of the device through the Learning Function. Information on constraints provided by the constraints knowledge was included in these clusters for recognition of constraints by the system.

For both controls, restrictions for controlling the pendulum and the cart exist in form of constraints. The control constraints were divided into two; the cart movement constraints and the pendulum rotation constraints.

The cart movement constraints are restrictions to the horizontal movements of the cart as shown in Figure 2.7a. The cart movements are limited due to these constraints. The pendulum rotation constraints are restrictions to the rotary movements of the pendulum as shown in Figure 2.7b. The pendulum rotation is restricted to a certain angle at a certain cart position due to these constraints. Due to the pendulum rotary movement being independence, the system must configure the pendulum rotary movement against these constraints using the cart movement indirectly.



(a) Constraints restricting the movement of the cart.

(b) Constraints restricting the rotary movement of the pendulum.

Figure 2.7: Constraints of the cart and the pendulum.

2.2.3 Learning Function for Substitute Target System

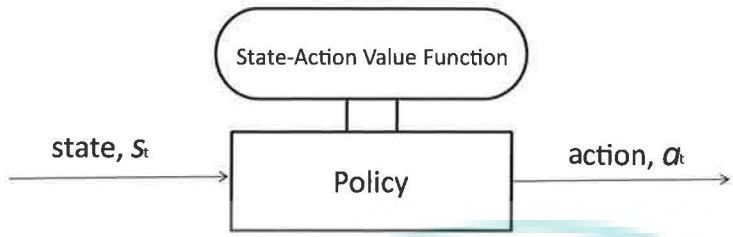
The Learning Function provides updates to the substitute target knowledge $Q(s, \Delta x)$ based on the state clusters assigned in recognition function. The update occurs during the downwards state of the pendulum, before the control function selects the substitute target displacement Δx for the cart movement controls. Reward defines the goal in reinforcement learning based on state clusters that determines reward r at a precise moment based on the state clusters shown in table 2.5.

Table 2.1: Reward settings for assigned state clusters.

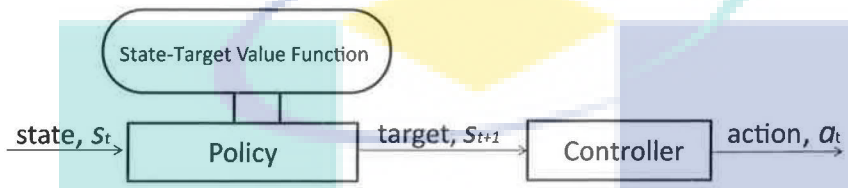
State Cluster	Reward, r
Near Control Objective State	$+r$
Over speed and Exceed Control Objective State	$-r$
Increasing Pendulum Swing Angle	$+r(^{\Delta\theta}/_{\pi})$
Decreasing Pendulum Swing Angle	0
Constraints Encounter	$-r$

*$\Delta\theta$ is the pendulum angular displacement from the initial position.

In order to provide a substitute target based learning agent into a control system, the Q-Learning algorithm introduced in 1.1 was modified for applying a value function that is based on these substitute targets. The target state is the expecting state s_{t+1} as reaction to action a_t. The value function does not defines target state s_{t+1} as action a_t; instead the target state displacement Δs_{t+1} from the current state s_t will defines the action required to achieve the target state s_{t+1}. Here, the distance towards the future state will replaces the action part of the conventional Q-learning into equation 1.1.



(a) Conventional reinforcement learning applies an action from a state action value function.



(b) Policy selects a target from a state-target value function that determines action.

Figure 2.8: State and action relation for substitute target based Q-learning.

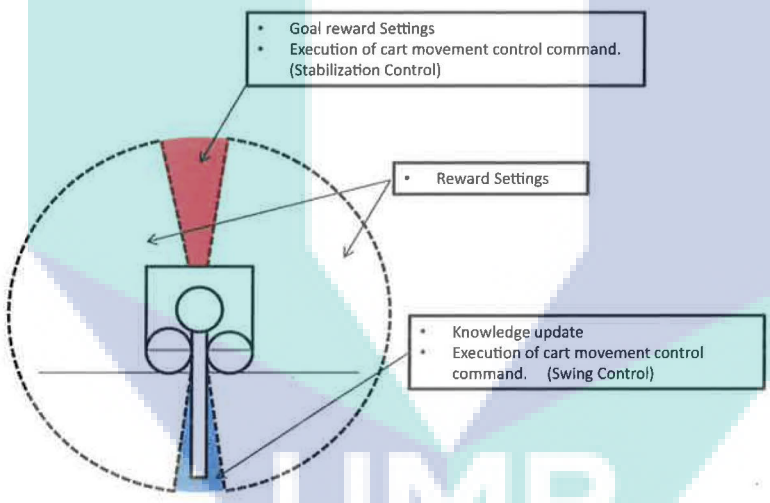


Figure 2.9: Control processes assigned according to pendulum angle.

$$Q(s_t, \Delta s_{t+1}) \leftarrow (1-\alpha)Q(s_t, \Delta s_{t+1}) + \alpha[r + \gamma \max_{\Delta s_{t+2}} Q(s_{t+1}, \Delta s_{t+2})], \tag{2.2}$$

Figure 2.8 explains the differences between the conventional Q-learning introduce in chapter 1. The relation of the state and action in conventional Q-learning utilizes state-action value function. Here, the action a_t is defined by a controller based on target state displacement Δs_{t+1} decided by the policy from state-target value function.

2.3 Experiments & Results

The effectiveness of the Learning Control System that utilizes substitute targets was confirmed through a series of simulations and a real machine test. The simulations were conducted to confirm the flexibility of the system to multi-functionality in conducted the swing up-control while avoiding the surrounding obstacles. The simulations started with confirmation of the effectiveness of the learning process, continues with confirmation of adaptability with direct constraints and indirect constraints. Results provided through these simulations should confirm the effectiveness of the Learning Control System in applying multi-functionality in such cases.

2.3.1 Experiments Settings

Due to application on the cart-pendulum device, a study on the parameters of the control device was done prior to constructing the simulations. The details according parameters involved in cart-pendulum device were analysed and prepared according to the diagram shown in figure 2.10.

2.3.1.1 Inverted Pendulum Model

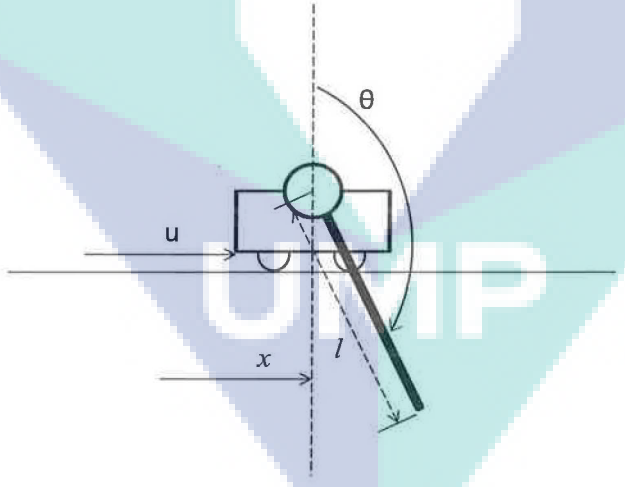


Figure 2.10: Diagram of the cart-pendulum parameters.

The mathematical model of the cart-pendulum device is derived to be applied in the simulation. Applying Newton's Second Law at the centre of gravity of the pendulum, the horizontal, X and vertical, Y components, are represented by

Table 2.2: Parameters description for cart-pendulum device.

Cart mass	M [kg]
Cart position	x [m]
Horizontal force on cart	u $[kgms^{-2}]$
Pendulum mass	m [kg]
Pendulum length	l [m]
Center of gravity to pivot length	L [m]
Gravitational acceleration	g $[ms^{-2}]$
Angular displacement	θ [rad]
Pendulum friction coefficient	c
Cart friction coefficient	d
Moment of inertia of the pendulum	I $[kgm^2]$

$$Y - mg = m\frac{d^2}{dt^2}(L\cos\theta) \tag{2.3}$$

$$X = m\frac{d^2}{dt^2}(x + L\sin\theta) \tag{2.4}$$

Both equations provides the torque equation,

$$I\ddot{\theta} + c\dot{\theta} = YL\sin\theta - XL\cos\theta \tag{2.5}$$

Applying Newton's Second Law to the above equation yields

$$u - X = M\ddot{x} + d\dot{x} \tag{2.6}$$

By substituting equations 2.3 and 2.4 into equations 2.5 and 2.6, the non-linear mathematical model of the cart-pendulum system can be derived as

$$\ddot{\theta} = \frac{1}{I + L^2m}[Lm(g\sin\theta - \ddot{x}\cos\theta) - c\dot{\theta}] \tag{2.7}$$

$$\ddot{x} = \frac{1}{M + m}[u - Lm(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta) - d\dot{x}] \tag{2.8}$$

The pendulum state of inverted position corresponds to an unstable equilibrium point $(\theta, \dot{\theta}) = (0, 0)$. In the neighbourhood of this equilibrium point, both θ and $\dot{\theta}$ are very small. Therefore, small angles of θ and $\dot{\theta}$: $sin(\theta) \approx \theta$, $cos(\theta) \approx 1$ and $(\dot{\theta})^2\theta \approx 0$. Thus, equation 2.7 and 2.8 can be rewritten as

$$\ddot{\theta} = \frac{1}{I + L^2 m}[Lm(g\theta - \ddot{x}) - c\dot{\theta}] \tag{2.9}$$

$$\ddot{x} = \frac{1}{M + m}[u - Lm\ddot{\theta} - d\dot{x}] \tag{2.10}$$

For the above two equations to be in a valid state matrix, $\ddot{x}$ and $\ddot{\theta}$ must be functions of lower order terms. $\ddot{x}$ and $\ddot{\theta}$ is substituted in equation 2.9 and 2.10, the state model is obtained as

$$\dot{s} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -kb & \frac{-(Lm)^2 gb}{I+L^2 m} & \frac{Lmcb}{I+L^2 m} \\ 0 & 0 & 0 & 1 \\ 0 & \frac{Lmka}{M+m} & Lmga & -ca \end{bmatrix} s + \begin{bmatrix} 0 \\ b \\ 0 \\ \frac{-Lma}{M+m} \end{bmatrix} u \tag{2.11}$$

$$y = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} s \tag{2.12}$$

where

$$a = \frac{M + m}{I(M + m) + L^2 mM} \tag{2.13}$$

$$b = \frac{I + L^2 m}{I(M + m) + L^2 mM} \tag{2.14}$$

Thus, the state and output vectors is represented by

$$s = \begin{bmatrix} x & \dot{x} & \theta & \dot{\theta} \end{bmatrix}^T \tag{2.15}$$

$$y = \begin{bmatrix} x & \theta \end{bmatrix}^T \tag{2.16}$$

2.3.1.2 Parameters for Control System Implementation

Parameters of the control object were selected based on a real cart pendulum device as shown in figure 2.11 prior to the simulation. Since the cart movement is limited to a certain range, the Learning Control System was applied on simulations before being handled by the real device. The simulations are programmed and arranged using MATLAB, based on the parameters of the real operating devices as shown in table 2.3.

Table 2.3: Parameters of the cart-pendulum device.

Cart Position, x	x [m]	
Cart Velocity, $\dot{x}$	v [m]	
Pendulum Angle, θ	θ [rad]	
Pendulum Angular Velocity, $\dot{\theta}$	ω [rad/sec]	
Cart mass, M	3.117 $[kg]$	
Pendulum Length, l	0.4 $[m]$	
Pendulum Mass, m	0.08 $[kg]$	
Pendulum Inertia, I	$\begin{bmatrix} 0.0012 & 0 & 0 \\ 0 & 1.6 \times 10^{-7} & 0 \\ 0 & 0 & 0.0012 \end{bmatrix}$	$[kgm^2]$

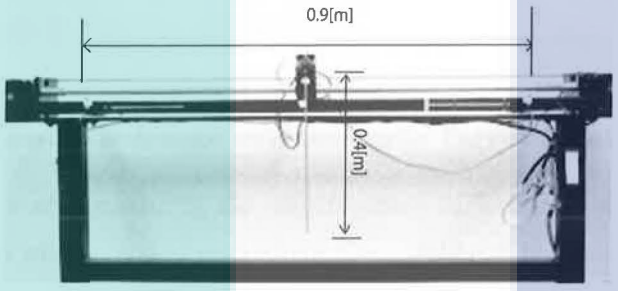


Figure 2.11: Cart-pendulum device (Japan E.M. Co., Ltd.) on which the simulations were based.

The simulations were conducted in three subjects according to the purposes; to confirm the effectiveness of learning using substitute target, to confirm the effectiveness of learning among direct control constraints and the effectiveness of learning among direct and indirect control constraints. Simulation for each subjects apply the parameters shown in table 2.4 for Q-learning which were selected prior to the experiments.

Table 2.4: Parameters for Q-learning of Learning Control System by substitute targets.

	Parameters	Range	Intervals
State	Cart Position, x[m]	-1.0 $\sim$ 1.0	0.2
	Pendulum Angular Velocity, ω [rad/s]	-14 $\sim$ 14	2
Substitute Target Displacement	Cart Movement Displacement, Δx [m]	-0.2 $\sim$ 0.2	0.05

Learning rate, α	0.5	Discount rate, γ	0.3

The initial and target states of the control device that was conducted in the simulation are as shown in table 2.5.2. During the initial state, the pendulum angle $\theta = \pi$ [rad], the cart position x are located in the middle of the track, $x = 0$ [m], while the cart velocity v and the pendulum angular velocity ω are zero. The inverted state are defined as target state, where the pendulum angle θ is 0 [rad]. The cart position of the assigned target state is the final substitute target x_T selected during the swing-up process.

Table 2.5: Initial state and target state of the simulations for Learning Control System by substitute target.

Parameters	Initial State	Target State
Cart Position, x[m]	0	x_T
Pendulum Angle, θ[rad]	π	0
Cart Velocity, v[m/s]	0	0
Pendulum Angular Velocity, ω[rad/sec]	0	0

2.3.1.3 Simulation on the Usage of Substitute Target Knowledge

In case of the subject of confirming the effectiveness of learning using substitute target knowledge, simulation was arranged to confirm the validity of the Learning Control System using constructed substitute target knowledge and a random substitute target knowledge.

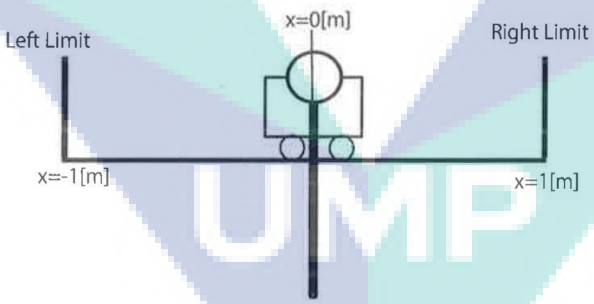
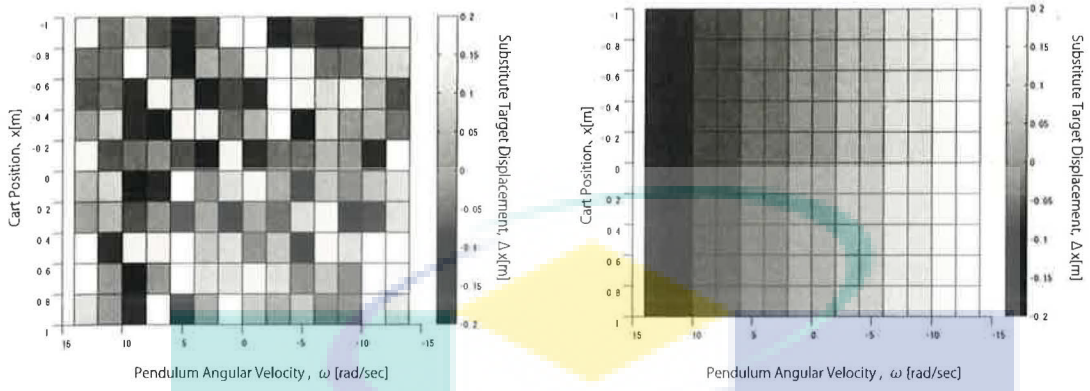


Figure 2.12: Movement range of the cart to satisfy the knowledge limit.

The constructed substitute target knowledge, as shown as figure 2.17b, was structured based on the basics of pendulum swing intensification control without arrangement of any constraints. The random substitute target knowledge as shown in figure 2.13a was structured as the value function by random number.

For this subject, the simulation was conducted in episodes, the simulation stops after



(a) Random control knowledge. (b) Constructed control knowledge.

Figure 2.13: The substitute target knowledge used as initial state.

each trial which counts as one episode. The rules assigned for the simulation were specified into:

- Simulation completes after 2300 episodes.
- Policy selects substitute target by *roulette* selection for 2000 episodes.
- Policy selects substitute target by *greedy* selection for 300 episodes starting after 2000th episode.
- Stop the simulation for each episodes.
- If the pendulum is in the inverted position, episodes end.
- If the cart position is out of learning range, episode ends

The *roulette* selection is a selection policy assigned to help increase exploration rate of a value function by turning the value of its selection options into selection probability and selected based on a random number. The *greedy* selection is a selection policy that selects an optimum option from the value function based on the highest value.

2.3.1.4 Simulations on Learning Control through Direct Control Constraints

These simulations were arranged for confirming the capability for learning control through direct control constraints, which is the constraints within cart movement, x. For this subject, three simulations were performed with three different sets of constraints as shown in figure 2.14. The constraints of the cart position x as shown in Figure 2.14 are as follows:

(a) Case 1 (P8M8), $-0.8 < x < 0.8$: Constraints on both left and right sides

(b) Case 2 (P4M4), $-0.4 < x < 0.4$: Larger constraints on both left and right sides

(c) Case 3 (P2M6), $-0.6 < x < 0.2$: Constrains on the right side is larger than that on the left side.

Constructed substitute target knowledge as shown in figure 2.17b were used as the initial knowledge fore these simulations in hope of a shorter simulation time. In order of

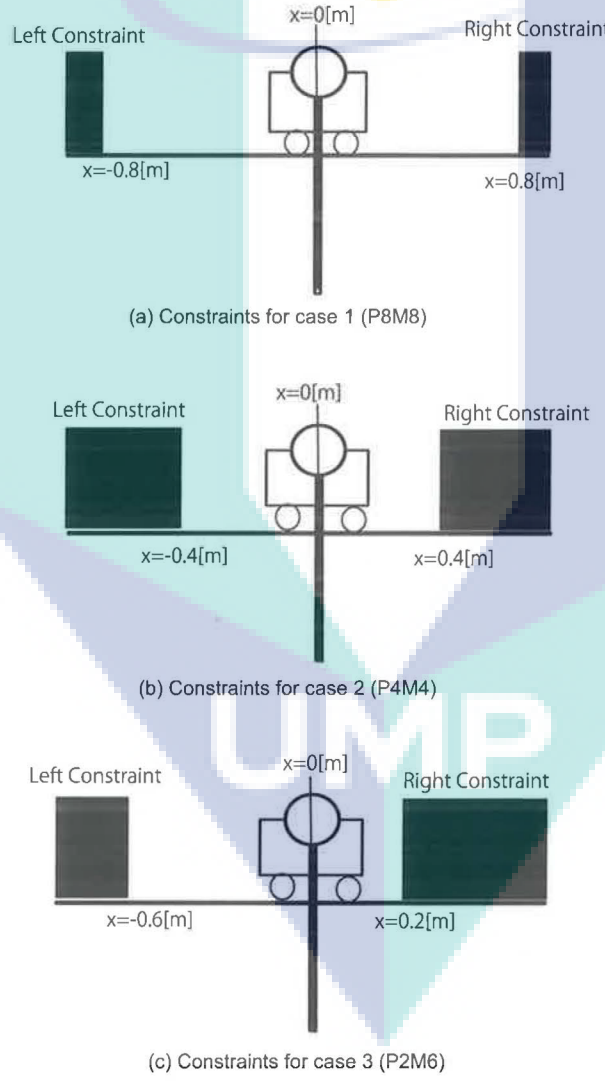


Figure 2.14: Constraints assigned among the cart position for simulations.

real control object application, the required behaviours for the system control object were specified before the simulation was done:

- Stop the simulation after 3000 seconds.
- If the pendulum is in the downward position, initialize after 10 times knowledge renewal.
- If the pendulum is in the inverted position, initialize after 3 seconds.
- If the cart position is between constraints, initialize after encountering the constraints.
- Attenuate the pendulum swing during initialization.

2.3.1.5 Simulations on Learning Control through Direct and Indirect Control Constraints

These simulations was arranged for confirming the capability of the proposed system for learning through indirect control constraints, which is the constraints in the pendulum angle θ [rad]. For this subject, three simulations were conducted consisting both direct and indirect control constraints. These simulations were categorized into 3 cases which each have difference sets of constraints.

(a) Case 1: Cart movement constraints in left and right side

(b) Case 2: Cart movement constraints in left and right side, and pendulum rotation constraints at top left and top right side.

(c) Case 3: Cart movement constraints in left and right side, and pendulum rotation constraints at the top middle.

During simulation, substitute target displacement were selected from the knowledge based on figure 2.17b by *roulette* selection for 300 trials, and later continues with *greedy* selection.

The desired system behaviour during simulation is described below.

- Stop the simulation after 5000 seconds.
- If pendulum is in downwards position, initialize after 25 times knowledge renewal.

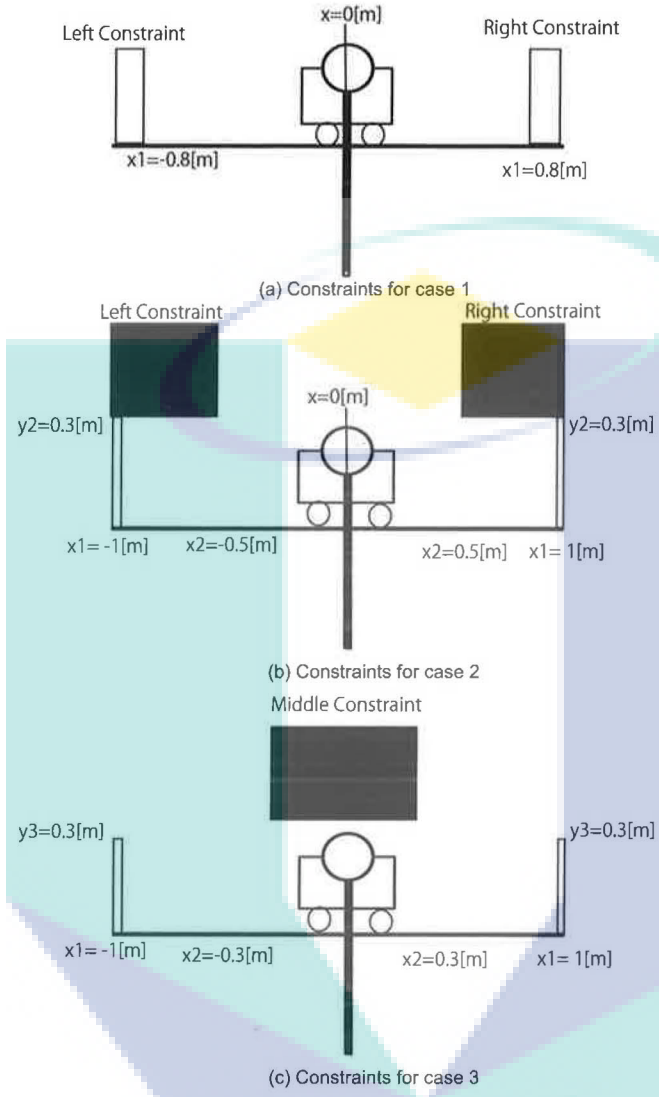


Figure 2.15: Constraints arranged around the cart position and the pendulum angle.

- If the pendulum is in inverted position, initialize after 3 second.
- If cart position is between constraints, initialize after constraints encounter.
- If pendulum tip point is inside constraints area, initialize after constraints encounter.
- Attenuate the pendulum swing during initialization.

2.3.2 Experiments Results

Results for application of substitute target on Learning Control System emphasize the effectiveness of the system into providing multi-functionality in a cart-pendulum system. The Simulations provide results concerning the effectiveness of the system in developing substitute target knowledge. Results from the simulations were then applied on the real cart-pendulum device to confirm the effectiveness of the system in real world application.

2.3.2.1 Experiments Results by Simulations

Results from simulations are separated in three parts according to three subjects arranged in the settings of the simulation. Firstly, results concerning the effectiveness in developing the substitute target knowledge were analysed. Then, results concerning development of substitute target knowledge by controls among direct constraints were analysed. Finally, results concerning development of substitute target knowledge by controls among direct and indirect constraints were analysed,

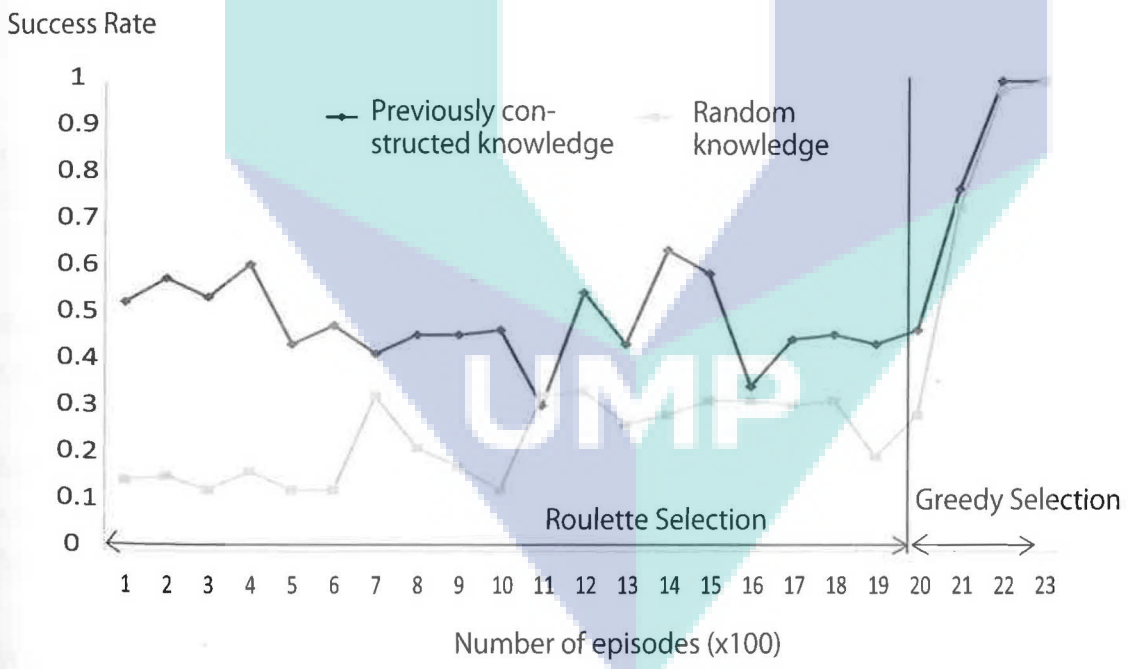
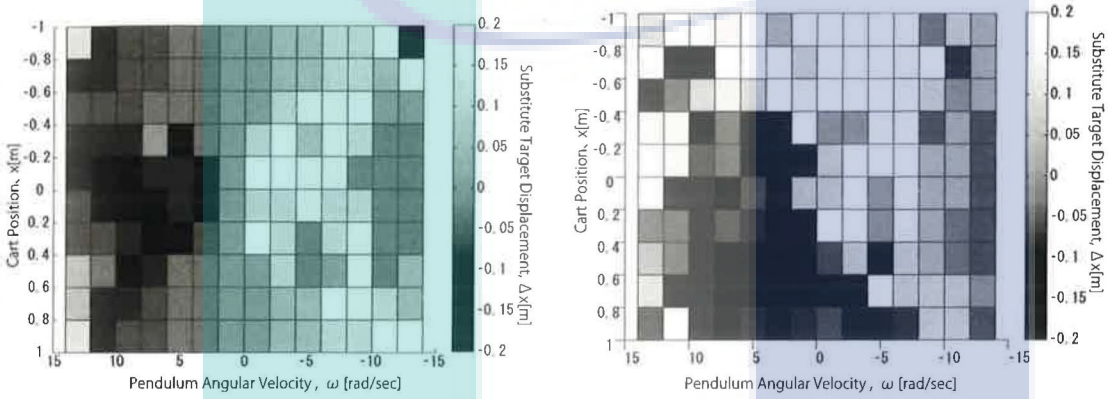


Figure 2.16: The average control success rate of the swing control for the first subject of the simulation.

2.3.2.1.1 Effectiveness of the Developing Substitute Target Knowledge.

Figure 2.16 provides information about the development of the substitute target knowledge according to episodes during the simulation. Here, the result shows that the rate of successful episodes increases towards maximum at the end of the simulation. Simulation using constructed knowledge started at a higher successful rate compared to simulation using random knowledge. The developed knowledge is shown in figure 2.17, describing that updates were applied on the knowledge, changing the structure of the value functions during the simulation.



(a) Updated constructed knowledge.

(b) Updated random knowledge.

Figure 2.17: The updated knowledge for both constructed knowledge and random knowledge after the simulation for Learning Control System by substitute targets.

2.3.2.1.2 Development of Substitute Target Knowledge from Control among Direct Constraints

Figure 2.18 provides information about the development of the substitute target knowledge according to episodes during the simulation with assigned direct constraints along the cart movement path. Here, the result shows that the rate of successful episodes increases towards maximum at the end of the simulation for all three cases of direct constraints.

The cart movement manoeuvre that was obtained in the substitute target knowledge is shown in 2.19, showing that the cart-pendulum system was able to successfully avoided the assigned constraint states. Each movement successfully swung the pendulum towards the inverted state using the path available for moving the cart. Here, results shows that the substitute target knowledge is able to constructed safe and reliable control knowledge

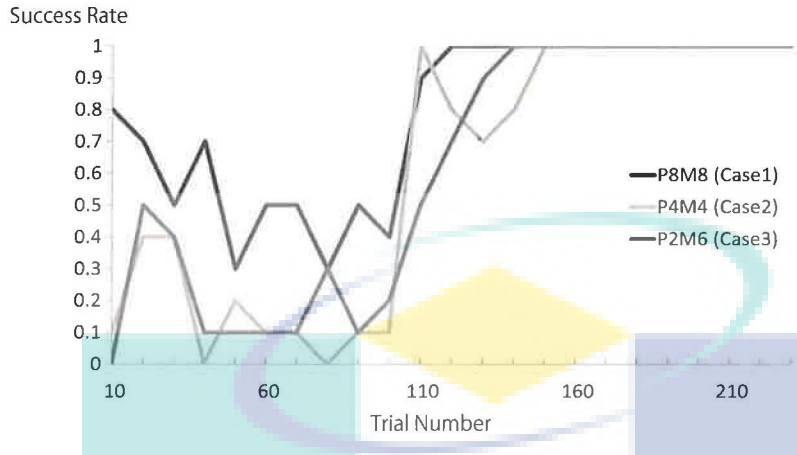
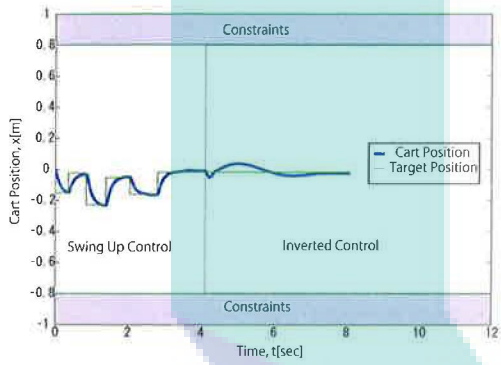
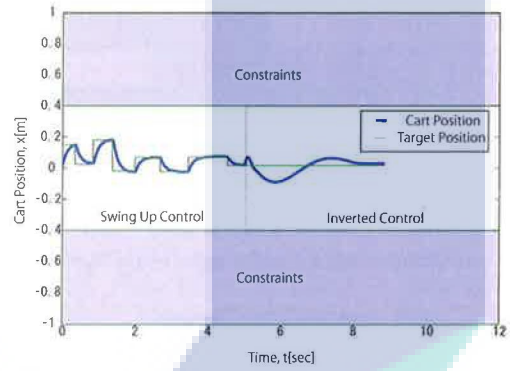


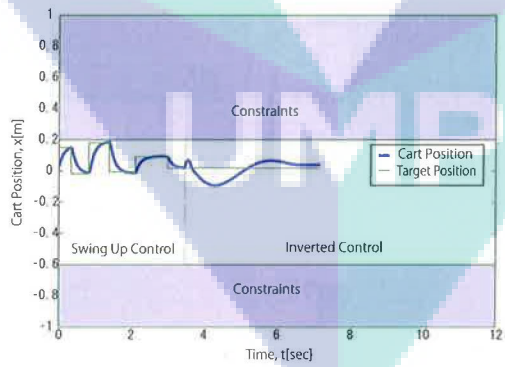
Figure 2.18: The average success rate for pendulum swing control among direct control constraints.



(a) Cart movements during P8M8 simulation.



(b) Cart movements during P4M4 simulation.

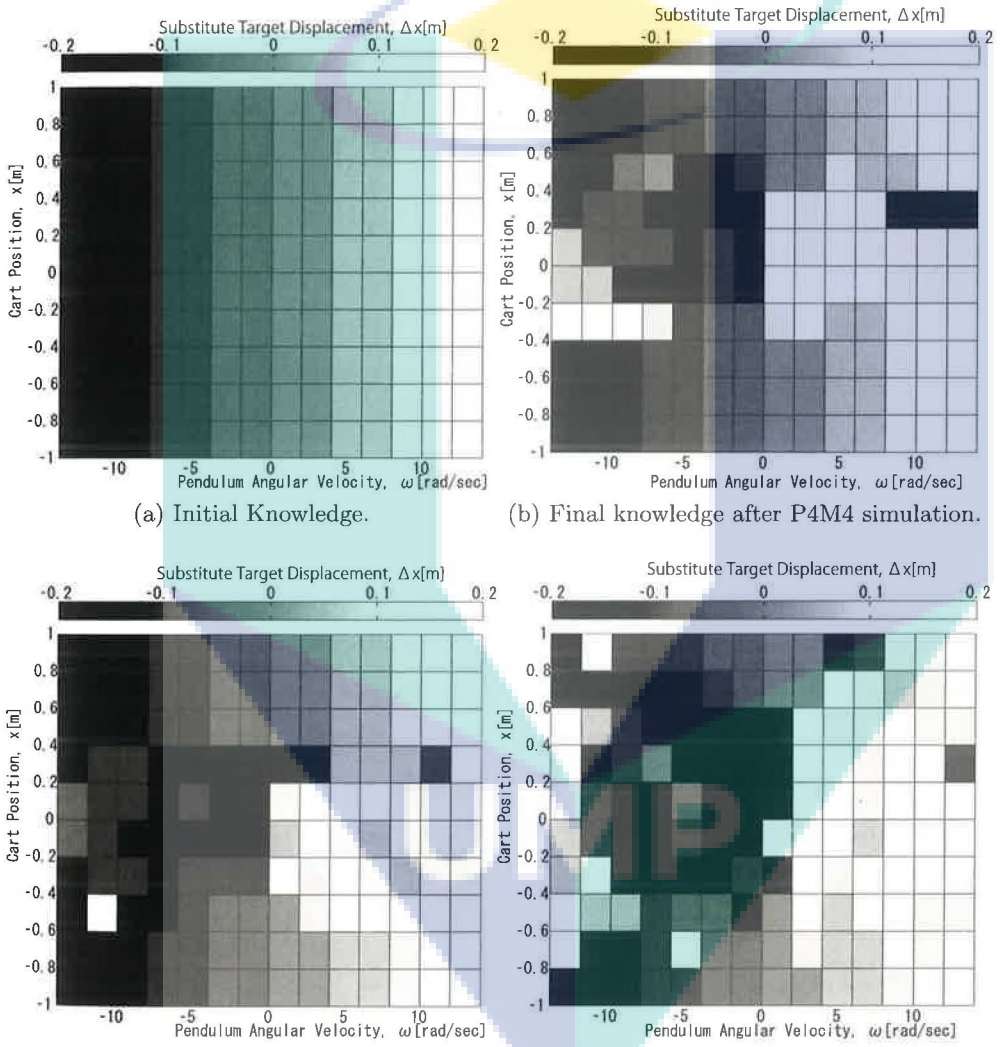


(c) Cart movements during P2M6 simulation.

Figure 2.19: Cart movement during the successful swing control for simulations for the three assigned cases of direct control constraints.

under influence of direct constraints. The control knowledge that applies these manoeuvres is shown 2.20.

The developed knowledge is shown in figure 2.20 describing that updates were applied on the knowledge when compared with the initial knowledge, changing the structure of the value functions during the simulation. Different structures were obtained due to the influence of constraints on the development of substitute target knowledge.



(a) Initial Knowledge. (b) Final knowledge after P4M4 simulation.

(c) Final knowledge after P2M6 simulation. (d) Final knowledge after P8M8 simulation.

Figure 2.20: Comparison between initial substitute target knowledge and final substitute knowledge for each simulation concerning the direct control constraints.

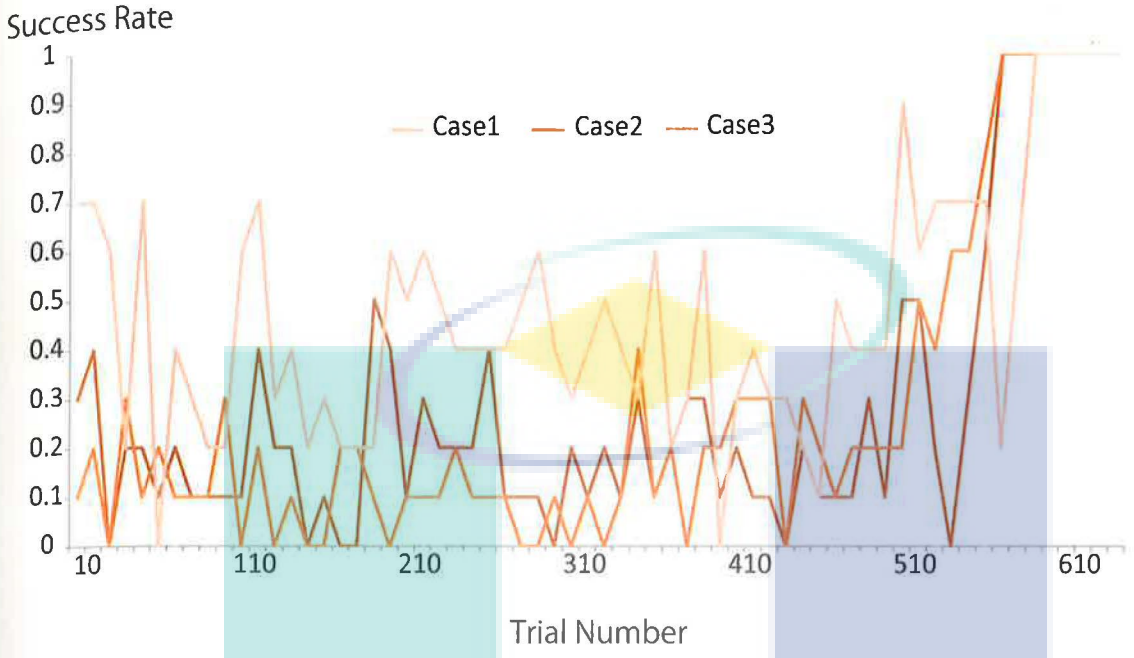


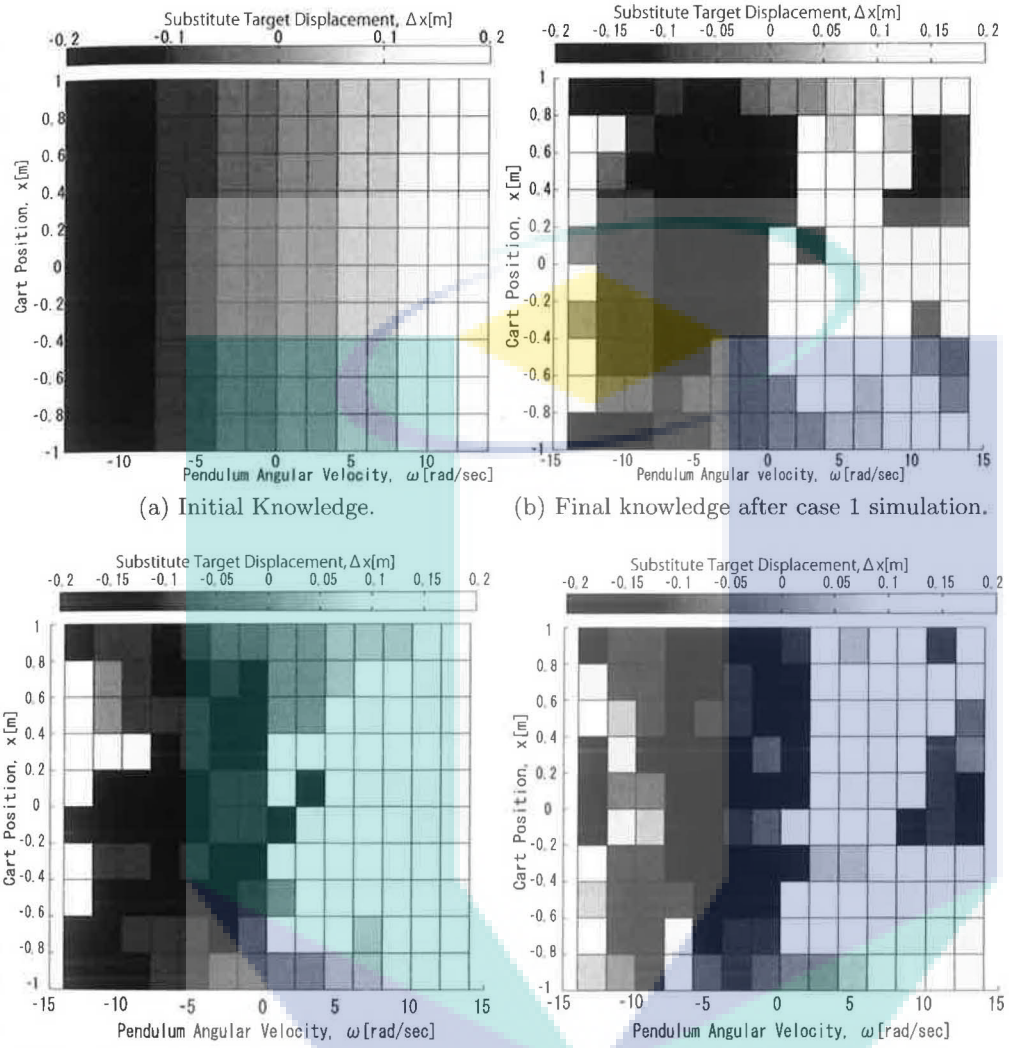
Figure 2.21: The average success rate from every 10 trials for Pendulum Swing control among direct and indirect control constraints.

2.3.2.1.3 Development of Substitute Target Knowledge from Control among Direct & Indirect Constraints.

Figure 2.21 provides information about the development of the substitute target knowledge according to episodes during the simulation with assigned direct constraints and indirect constraints along the cart and pendulum movement path. Here, the result shows that the rate of successful episodes increases towards maximum at the end of the simulation for all three cases of direct and indirect constraints.

The developed knowledge is shown in figure 2.23 describing that updates were applied on the knowledge when compared with the initial knowledge, changing the structure of the value functions during the simulation according to the assigned constraints. Different structures were obtained due to the influence of direct and indirect constraints on the development of substitute target knowledge.

The constraints that was detected by the Learning Control System by substitute target during this simulation is shown in 2.23 showing that the cart-pendulum system was able to successfully avoided the assigned constraint states. Results show that the constraints state can be detected and avoided by the Learning Control system by substitute targets. Control



(a) Initial Knowledge.

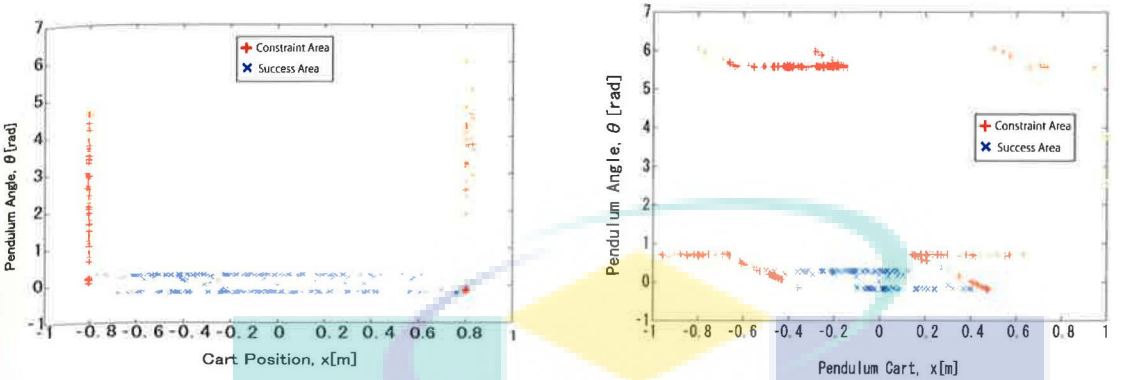
(b) Final knowledge after case 1 simulation.

(c) Final knowledge after case 2 simulation.

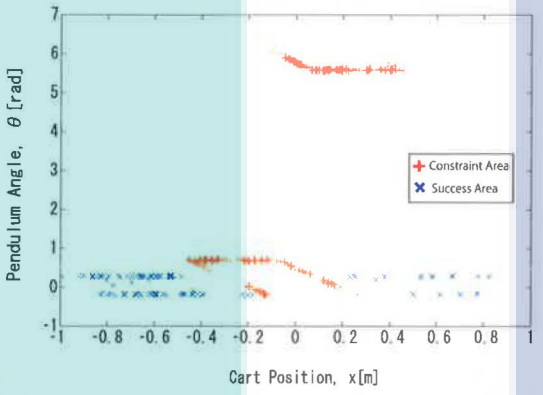
(d) Final knowledge after case 3 simulation.

Figure 2.22: Comparison between initial substitute target knowledge and final substitute knowledge concerning direct and indirect control constraints.

manoeuvre is configured by the substitute target knowledge that had experienced collision with the assigned constraints. The Learning Control System by substitute target are able to provide multi-functionality by being able to swing the pendulum towards inverted state while avoiding any assigned constraints. Here, results shows that the substitute target knowledge is able to constructed safe and reliable control knowledge under influence of direct and indirect constraints.



(a) Specific area detected during simulation for case 1. (b) Specific area detected during simulation for case 2.

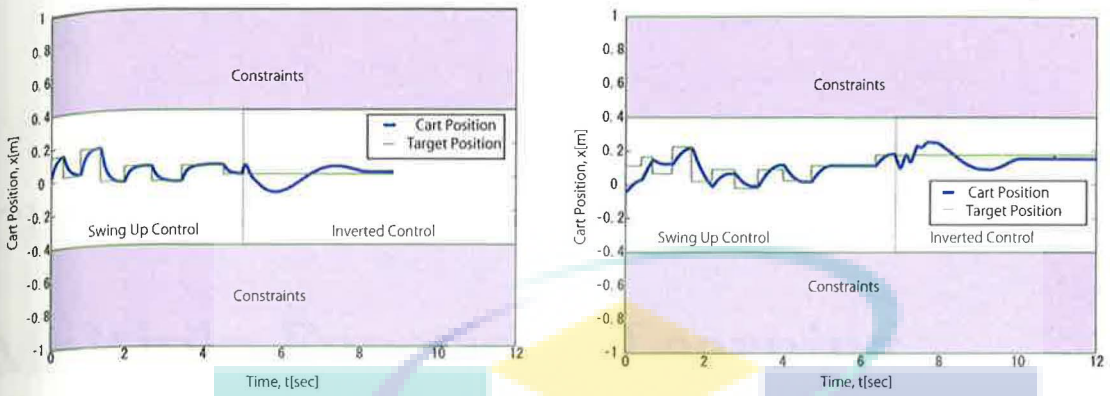


(c) Specific area detected during simulation for case 3.

Figure 2.23: Success area and constraint area detected during the simulations with direct and indirect constraints.

2.3.2.2 Experiments Results by Real Machine Operation

After confirming the effectiveness of the Learning Control System by substitute targets in the simulation, real machine test was conducted using results obtained in one of the simulation. Here, simulation of case P4M4 in simulation of subject 2 is applied due to the parameters being utilizable on the real cart-pendulum device. Comparison of the results obtained through simulation and real machine can be seen in figure 2.24. The control manoeuvre differs due to slight differences of the real device properties compared to the specification applied on the simulation. However, the real device was able to apply the substitute target knowledge in conducting a new safe control manoeuvre. Here, results show that the Learning Control System by substitute target is applicable in real operation.



(a) Cart movements during P4M4 simulation

(b) Cart movements from the application of P4M4 final knowledge on real machine

Figure 2.24: Cart movement during the successful swing control for simulation P4M4 and result from application on a real machine.

2.4 Summary

A learning control system that learns substitute target knowledge is designed to provide multi-functionality for a safe and reliable control in achieving the final target while considering constraints. Application of substitute target may provide utilization of linear control knowledge in a complex non-linear control system. Application of substitute target was utilized on cart-pendulum control where constraints were assigned in the cart and pendulum movement path. Simulations was arranged to confirm the effectiveness of the system in applying multi-functionality by providing successful swing controls among assigned constraints.

The Learning Control System was able to learn to consider environment constraints while learning to control the control device. During simulation, constraints were detected by the Learning Control System and the system learns to construct a safer control manoeuvre considering the assigned constraints. The substitute target knowledge learned in one of the simulation is applied on real control operation and results shows that the Learning Control System by substitute target are applicable on real world operation.

Based on the results, safe and reliable controls were obtained through utilization of substitute target in a Learning Control System. Applying substitute targets in a Learning Control system could provide multi-functionality, resulting in safer and reliable controls for non-linear devices.

Chapter 3

Multiple-Functions Learning Control by Multiple Control Knowledge

Design of Learning Control with quality of multi-functionality produces functions with each function are based on specific control knowledge. Utilizing multiple control knowledge in a Learning Control System may provide flexibility in producing control commands, where relevant control function can be chose according to the requirement of the control environment. Design of a Learning Control System that applies multiple control knowledge may provide human like multi-functionality where human dependency can be reduced, resulting in semi-autonomous control device.

3.1 Multiple Control Knowledge in Learning Control

Human command plays major role in providing instruction for a device through series of control systems. Such command is based on human decisions in monitoring the surrounding environment, choosing an optimum option in providing reliable manoeuver to the control device. Complex control system such as devices with non-linearity produces more strain in the human decisions, requiring expert skills in producing command for a safe and reliable control. Applying a Learning Control System with multiple control knowledge can help decide a control decision to support an operation and can reduce the dependency on human command through application of multiple source of control knowledge in the system. Multiple source of control knowledge can be updated using Learning Control, providing expert control capable of replacing human commands.

Multiple source of control knowledge provides multiple options of functions in a control System that has potential to produce human like multi-functionality in a control operation. This is due to autonomous development of multiple control knowledge provided by the Learning Control System as shown in figure 3.1. Using multiple source of control knowledge, control strategy that applies both control knowledge can be produced, resulting in expert control of the control device. Here, the Learning Control System with multiple control knowledge can provide most of the control decision, reducing the control burden on human command.

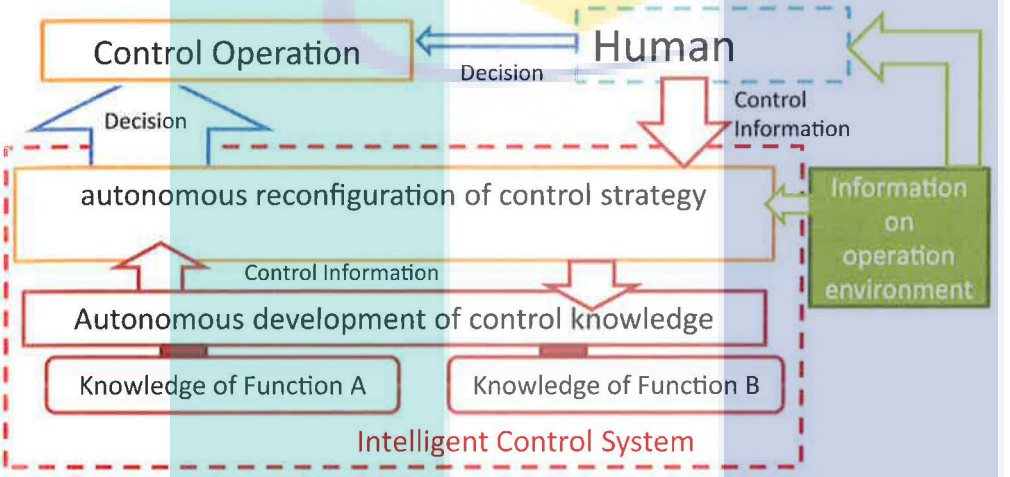


Figure 3.1: Structure of Learning Control System by multiple control knowledge.

Learning Control System by multiple control knowledge may reduce burden for controls on control devices with non-linearity. Control device as aerial hovering vehicle shown in figure 1.3 requires the operator to control the movement of the control device while maintain the stability of the device on air. Expert human operator is capable in manipulating those control parameters for rapid position transition of such device. Human multi-functionality provides commands on the angular orientation of the device using cyclic with assistant of thrust command that is also provided by the human. Based on the human multi-functionality, a Learning Control System with multiple source of Control Knowledge may provide Rapid Position Control by multiple Acceleration Control Functions in aerial hovering vehicle. Here, a Learning Control System by multiple source knowledge is design to provide rapid position control and rapid position control among obstacles for aerial hovering vehicle. The system is separated into two sections where the first section introduces the design of Learning Control System by multiple control knowledge for rapid position con-

trol, while the second section introduces the design of Learning Control System by multiple control knowledge for rapid position and obstacle control.

3.2 Application of Multiple Control Knowledge in Learning Control: Rapid Position Control

The first section of the Learning Control System by multiple control knowledge was designed for rapid position control of aerial hovering vehicles. Aerial hovering vehicles consist of non-linear parameters that require expertise in providing a quick reliable control. Here, human expertise in operating such device is generated through application of Learning Control System by multiple control knowledge.

3.2.1 Introduction to Rapid Position Control

Controls for aerial hovering vehicles involve manipulation of cyclic and thrust. Expert operator is able to operate the cyclic and thrust in providing safe position control for aerial hovering vehicle as shown in figure 1.3 through non-linear parameters within the device. Expert operator could even perform rapid position transition using cyclic and thrust along obstacle due to skills and experience in operating such device. Such skill is difficult to be operated by an autonomous control system. Here, Learning Control System by multiple control knowledge is designed to provide expertise in rapid position control for aerial hovering vehicles.

The system was developed for learning the best coordination of target angle θ_T that can perform a rapid position transition. Target angle θ_T provides changing in the direction of the thrust to create horizontal force that can create a horizontal movement while airborne. Figure 3.2 shows the changing in direction of the thrust according to target angle θ_T making the horizontal movement possible.

Configuration of the target angle θ_T requires increasing in thrust for providing lift force to preserve the leaning angle against gravity. When the preservation period of the leaning angle increased, the horizontal velocity of the aerial hovering vehicle will be increased due to changing of intensity in the horizontal force. Therefore, certain strategy concerning configuration of the target angle θ_T and its preservation period is needed for providing acceleration and deceleration for a precise position control.

Figure 3.3 shows the manipulation angular orientation of the aerial hovering vehicle during a position transition. A target angle θ_T^1 is configured to provide a horizontal force

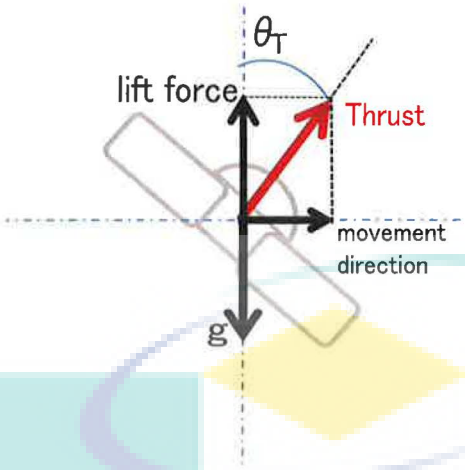


Figure 3.2: Configuration of aerial hovering vehicle by angular orientation.

for acceleration while another θ_T^2, is configured to provide a horizontal force for deceleration before returning to its initial angle θ_0. Such manipulation of target angles provides position transition between two point of x. Manipulation of target angles and thrust provide quick position transition which defined here as rapid position control.

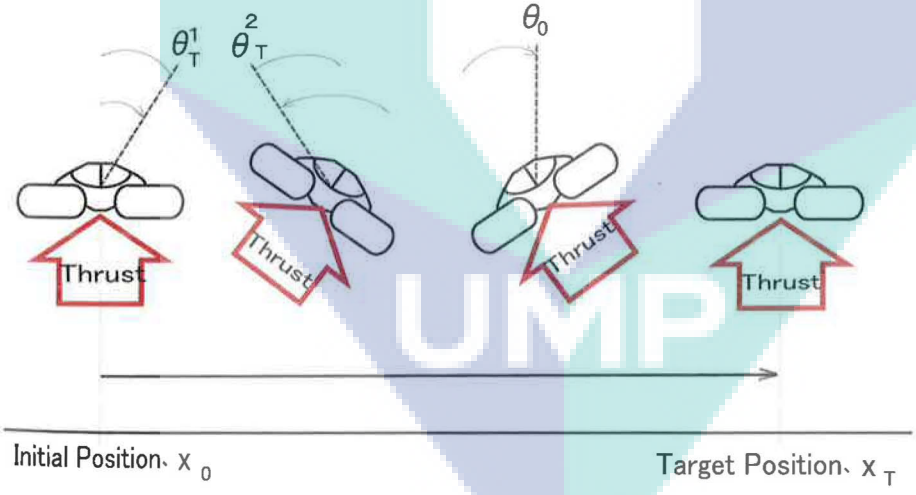


Figure 3.3: Position control of an aerial hovering vehicle using target angle θ_T as reference.

3.2.2 Manipulation of Angular Orientation for Rapid Position Control

Dynamics of the aerial hovering vehicles provides information concerning parameters that involves in creating rapid position control on such device. Using this information, controls of aerial hovering vehicles was emulated on cart-pendulum system, where a Learning Control System by multiple control knowledge was designed.

3.2.2.1 Emulating Angular Orientation Control of an Aerial Hovering Vehicle on Pendulum System

Controls of the aerial hovering vehicles are based on the non-linear properties of the device. Such properties can be defined in other control devices such as the cart-pendulum device. The dynamics of the aerial hovering vehicles concerning manipulation of angular orientation and thrust was emulated in the inverted control of cart-pendulum system as shown in figure 3.4. Manipulation of angular orientation is manipulated through manipulation of leaning angle of the pendulum while the horizontal force from the thrust was emulated through the cart movement. Through manipulation of pendulum's leaning angle and cart's movement, a movement similar to aerial hovering vehicle can be obtain where manipulation of the learning angle provides horizontal force, performed by cart movement, for applying position transition to the cart. Figure 3.5 shows the position control of cart-pendulum system which is emulated from the position control of the aerial hovering vehicles shown in figure 3.3.

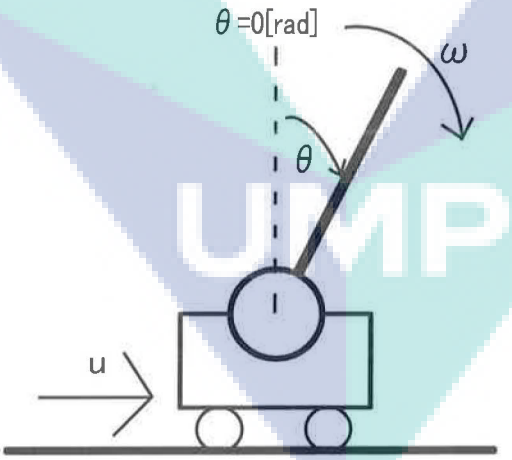


Figure 3.4: The stabilization control of inverted pendulum.

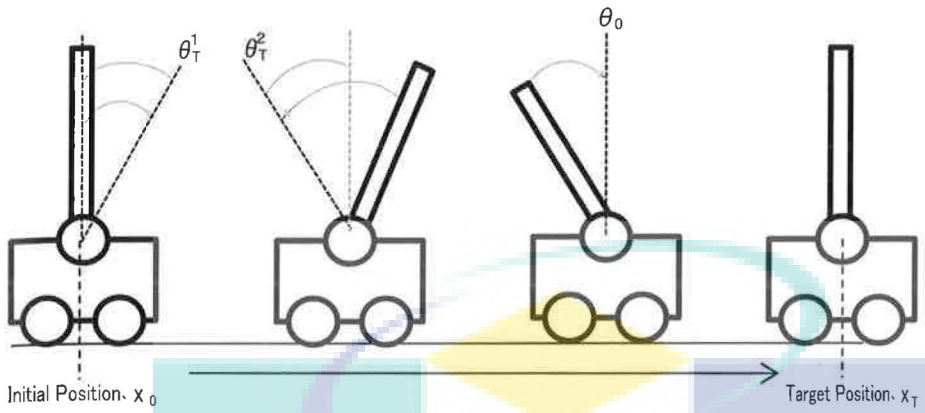


Figure 3.5: The position control of inverted pendulum using target angle θ_T as reference.

The above control of cart-pendulum was designed through Learning Control System by multiple control knowledge as shown in figure 3.6. Two control knowledge concerning manipulation of the leaning angle of the pendulum was embedded in the system, where the first control knowledge is about preservation time of learning angle for acceleration while the second control knowledge is the knowledge of preservation time of leaning angle for deceleration. Target learning angle was determined depending on the target cart location, therefore, combination of optimum leaning angle and preservation time of the angle has to be learned by the system to provide an optimum rapid position control of the cart-pendulum device. Here, the Learning Control System by multiple control knowledge is able to learn to provide expert control of the device, which emphasizes the possibility of applying such system on aerial hovering vehicle.

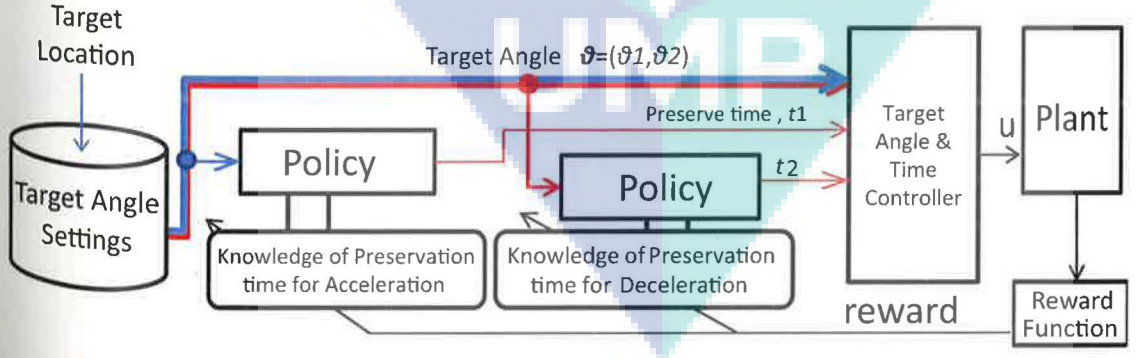


Figure 3.6: Structure of system with multiple control knowledge for rapid position control.

3.2.2.2 PD Control of Angular Orientation on a Cart Pendulum System

In order to apply the Learning Control System by multiple control knowledge on the cart-pendulum device, a series of test is done to confirm that the cart-pendulum device are capable of operating by using target angle θ_T as reference for the cart movement. It is known that the horizontal acceleration of aerial hovering vehicles increase when the leaning angle increased, therefore the same conclusion must be confirmed in the cart-pendulum device before being applied in the experiment. Here, a PD control system that applies target angle θ_T as reference for the cart movement was structured for the cart-pendulum device. Figure 3.7 shows the structure of the PD control of the cart-pendulum device using target angle θ_T as reference for the control command of cart movement.

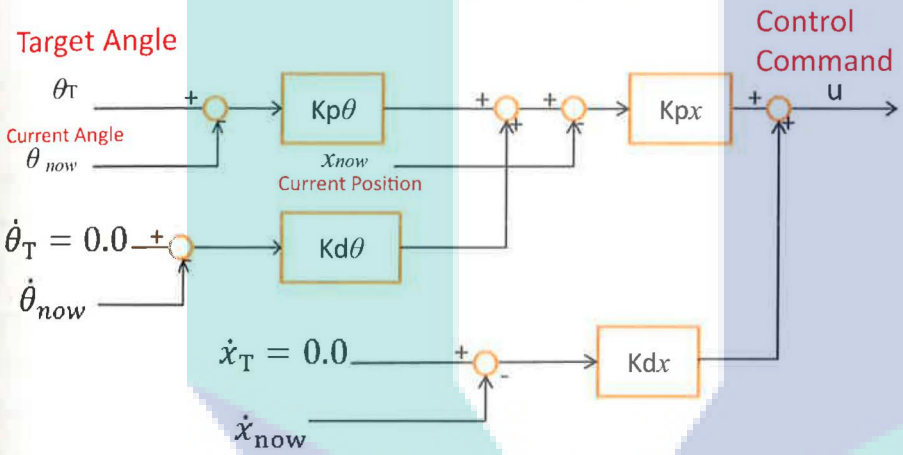


Figure 3.7: PD control of cart-pendulum system emulating aerial hovering vehicle.

Simulations were conducted to confirm the effectiveness of the structured PD control. Simulation was conducted using a set of target angle with arranged preservation time and the movement of the cart was recorded. Table 3.1 provides the results of the simulation that uses a set of three target angle θ_T. Here, it is confirmed that acceleration of cart movement increases as target angle θ_T increase. Control output concerning the motor input of the cart is studied to study the relation between the quantities of the output to the preservation period of the leaning angle. Figure 3.8 provides relation of the motor input and preservation time at every sampling pulse for each tested target angle θ_T.

Based on figure 3.8, a certain amount of output, total output V_{out} is produced for each target angle θ_T assigned for the leaning angle at a certain operation time. Total output V_{out}

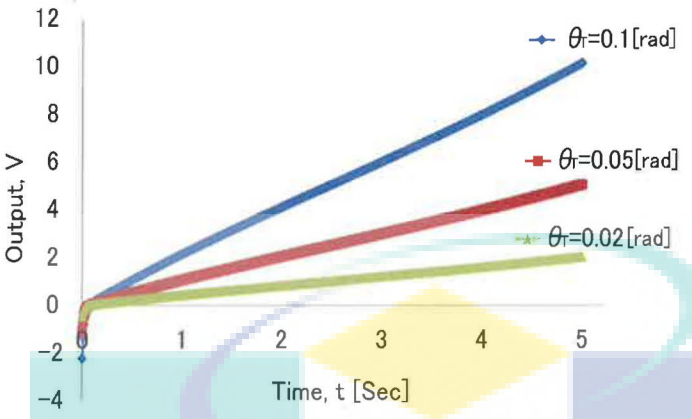


Figure 3.8: The reference data used to calculate the preservation period of output for each target angle.

provides information for calculating the period of maintaining the target angle, therefore, direct usage of Total output V_{out} is used in the learning algorithm to help produce the learning control system by multiple control knowledge. Total output V_{out} is used replacing the preservation time t as the unit of the preservation period of leaning angle is for lesser burden in computation during the simulation. Since the range of the cart movement for the operation is limited, the range of the total output V_{out} were limited up to 100 [V]. Based on these results, the angular orientation of the pendulum is known to be related to the acceleration of the cart movement. This confirmed that the system have the dynamics similar to the controls of an aerial hovering vehicle.

Table 3.1: Pre-experimental results for determining the output required by the cart-pendulum device for emulating the angular control of aerial hovering vehicles.

Target Angle,θ_T [rad]	0.02	0.05	0.1
Total output, V_{out}[V] needed to maintain θ_T for 3[sec] (Sampling time:0.01[sec])	184.5	460.0	919.6
Distant covered, x[m] in 3[sec]	0.80	2.00	4.02
Total output, V_{out}[V] needed to maintain θ_T for 5 [sec] (Sampling time:0.01[sec])	504.2	1259.8	2522.9
Distant covered, x[m] in 5[sec]	2.28	5.71	11.45
Average amount of output per distant covered, V_{out}[V/m]	225.89	225.28	224.55
Acceleration, a [ms^{-2}]	0.13	0.24	0.48

3.2.2.3 Obtaining Rapid Position Control Using Angular Orientation for Inverted Pendulum

Using the information obtained through the simulation for manipulation of angular orientation by cart-pendulum device, Learning Control System by multiple control knowledge for rapid position control of cart-pendulum device was designed as shown in figure 3.9. Here, Reinforcement Learning was applied for updating the control knowledge which in a form of value functions. The value functions consists of state and action parameters, where the state parameters is defined by target angle θ_T and action parameters is defined by total amount of control command output u for preserving target angle V_{out}. Target angle is defined by a set of setting rule, which based on command of target position assigned by a human operator. Target angle θ_T and total amount of control command output u is given to the controller for arranging control command u for the cart-pendulum device. Using the design of Learning Control System by multiple control knowledge for rapid position control, series of simulation was constructed for evaluating the effectiveness of the system.

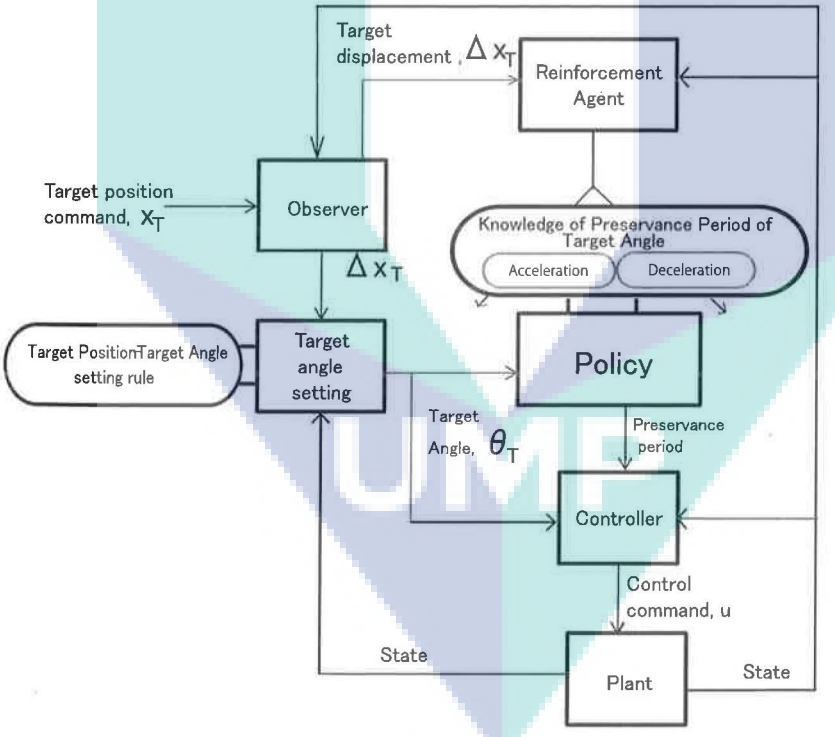


Figure 3.9: The structure of Learning Control System by multiple control knowledge for rapid position control of aerial hovering vehicles.

3.2.3 Simulation Settings

Experiment for confirming the effectiveness of the Learning Control System by multiple control knowledge in rapid position control was conducted in series of simulations. Q-learning was used to produce value functions $Q(\theta_T, V_{out})$ that defines the best combination of target angle θ_T and total output V_{out}. Q-learning algorithm updates the value functions $Q(\theta_T, V_{out})$ using reward r for producing an optimum control knowledge. The parameters of the Q-learning algorithm configured in the simulation are as shown in table 3.2, where these state and action parameters range was selected depending on the properties of the control device, selected prior to the experiment.

The algorithm is defined as

$$Q(\theta_T, V_{out}) = (1-\alpha)Q(\theta_T, V_{out}) + \alpha[r + \gamma Q_{max}], \tag{3.1}$$

$$Q_{max} = \max_{V'_{out}} Q(\theta'_T, V'_{out}), \tag{3.2}$$

where θ_T denotes continuing target angle θ_T and V_{out} denotes the total output V_{out} of the continuing target angle. α is denoted as learning rate while γ is denoted as discount rate.

The simulation was conducted by using five targets of cart position shown in figure 3.10. The objective of this simulation is to have the system learns the optimum control strategy for achieving the target cart position assigned by the multiple control knowledge assigned in the Learning Control System. All those targets were randomly selected before the simulations, where these five targets of cart position were selected to confirm that the system was able to learn a rapid position control at any direction and distance.

The properties and rules of the simulation were selected before conducting the simulation. These properties and rules are used for all five target positions assigned previously.

Table 3.2: Q-learning parameters for Learning Control System for rapid position control.

	Parameters	Range	Intervals
State	Target Angle, θ_T [rad]	-1 ~ 1	0.05
Action	Total Output, V_{out} [V]	0 ~ 100	20

Learning rate, α	0.5	Discount rate, γ	0.3

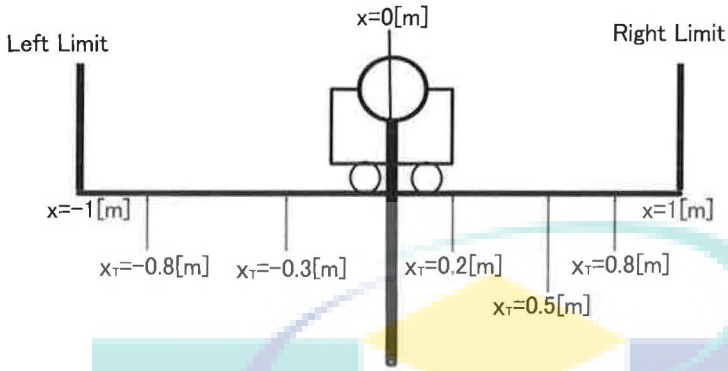


Figure 3.10: Target position assigned for simulation of rapid position control.

A control operation is defined by the process of attempting position control of the cart-pendulum device towards the target position within 10 seconds, where operation done is counted as trials. The other simulation properties are as follows:

- Simulation runs five times with different target position assigned.
- Simulation end at 550 trials.
- 10 seconds of operation time for each trial.
- $\epsilon-$greedy selection of output
- Reward is given after operation ends.
- Full reward, $r = 1$ is given to acceleration target angle, θ_T^1 if successfully achieve target position x_T at the end of an operation
- Half reward, $r = 0.5$ is given to deceleration target angle, θ_T^2 if successfully achieve target position x_T at the end of an operation
- zero reward, $r = 0$ is given to both target angle θ_T^1 and θ_T^2 if it fails to achieve target position x_T at the end of an operation.

The results were collected and analysed at the end of the simulation with 550 trials for each five assigned target position x_T conducted in the simulation.

3.2.4 Simulation Results

The result for the simulation was divided into two categories. The first category defines improvement achieved through the learning process while the second defines the successful control operation achieved at the end of the simulation. The improvement achieved in the first result confirms the validity of the learning process in creating a better control knowledge through the simulation that can lead to a successful control operation. The control operation shown in the second result confirms that the control operation operates the position transition towards the target position x_T successfully.

3.2.4.1 Knowledge Improvement through Learning Process

The value function $Q(\theta_T, V_{out})$ is at zeros at the beginning of the simulation, where any control operation operated under this control knowledge will less likely to be successful as no particular optimum combination of angle orientation can be detected from the knowledge. At the end of the simulation, the optimum combination is recognized through the update done by the Q-learning algorithm. Here, successful control operation is obtained and consistency is achieved in producing a successful position transition.

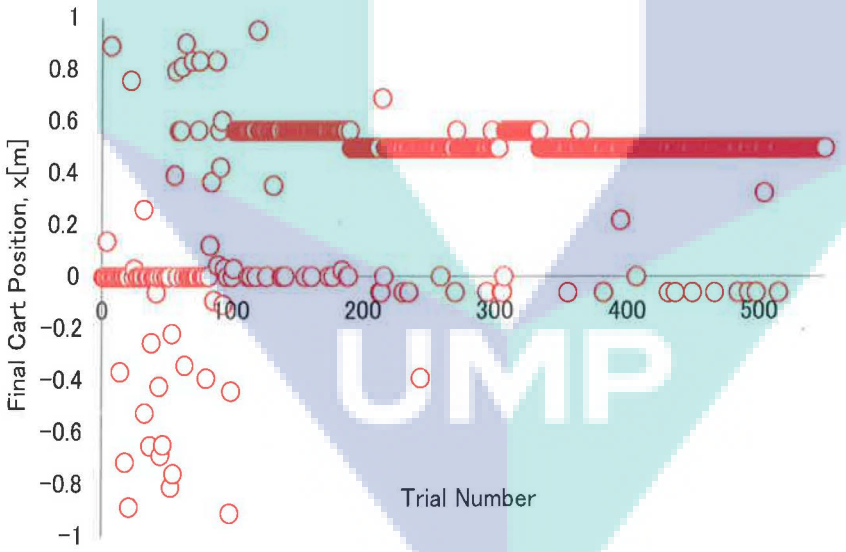


Figure 3.11: Improvement of the final cart position with respect to the number of trials.(Target position, x_T=0.5[m])

Figure 3.11 shows the results of position transition of the cart at the end of every operation trials. The results of cart positions at the beginning of the simulation are scattered

around the cart movement range. However, the results of the cart positions are focused to the target position at the end of the simulation. Here, a successful control operation that can achieve the target position is obtained.

3.2.4.2 Successful Operation Learned through Simulation

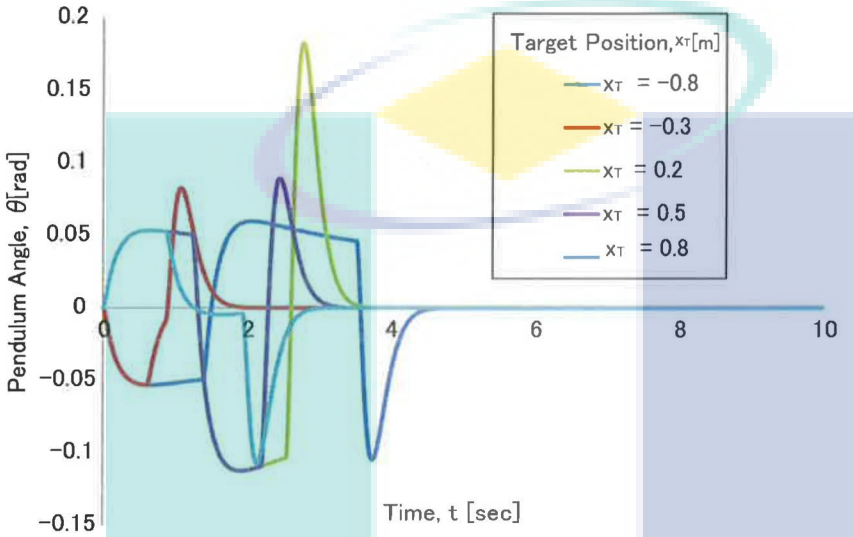


Figure 3.12: Angular trajectory of the pendulum during control operation that uses control knowledge obtained after 550 trials.

Figure 3.12 shows the pendulum angular trajectory during control operation that uses the control knowledge obtained after 550 trials. The pendulum trajectory during the control operation varies depending on each target position assigned. However, it can be seen that the pendulum angle stabilized at $\theta = 0[rad]$ around 5 seconds. Figure 3.13 shows the cart trajectory during control operation that uses control knowledge obtained after 550 trials. The cart trajectory is seen to be moving towards the target position and stabilizes near the target position with an error margin around ± 0.1[m].

The details of the successful control operation is shown in table 3.3. Here, for each cart position, specific acceleration angle θ_T^1 and deceleration angle θ_T^2 was selected to complete the control operation at certain amount of output V_{out}. The target angles θ_T^1 and θ_T^2 that were selected during the control operation provide a certain pattern. Acceleration angle θ_T^1 was leaning towards the direction of the target position x_T. However, deceleration angle were leaning to either the opposite direction of the target position x_T or zero. Here, the

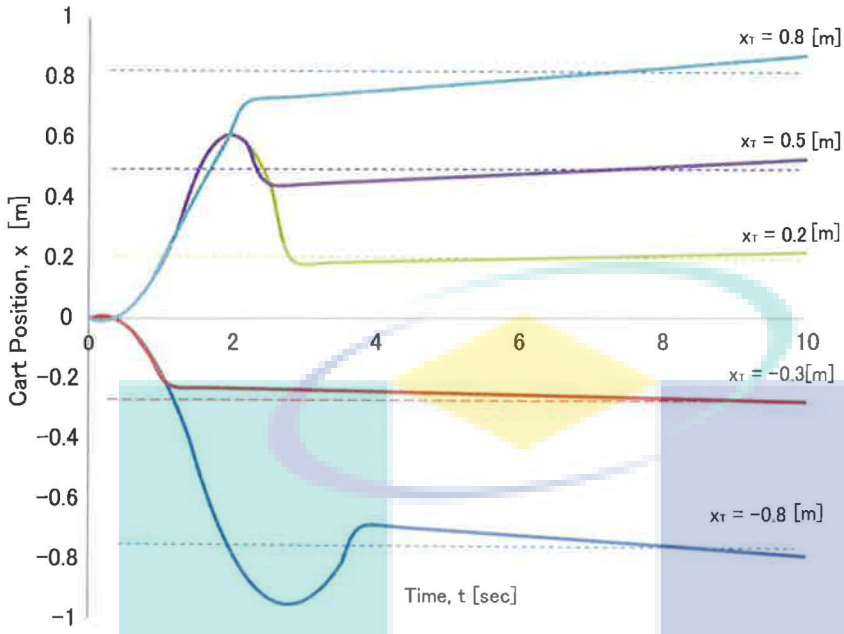


Figure 3.13: Movement trajectory of the cart during control operation that uses control knowledge obtained after 550 trials.

Table 3.3: Time required to complete a position control during a successful operation.

Target Position, x_T	-0.8	-0.3	0.2	0.5	0.8
Acceleration angle,θ_T^1 [rad]	-0.05	-0.05	0.05	0.05	0.05
Deceleration angle,θ_T^2 [rad]	0.05	0.05	-0.1	-0.1	0.0
Acceleration output,V_{out} [rad]	100	80	80	80	40
Deceleration output,V_{out} [rad]	80	40	60	20	100
Time until achieved stabilization, t[sec]	3.8	1.2	2.8	2.5	2.3

system learns that deceleration angle θ_T^2 was selected to decelerate for attempting to stop at the target position x_T. The total output V_{out} varies according to the target angle θ_T, depending on the required force for achieving the target position x_T.

Based on the result, the Learning Control System by multiple control knowledge was able to learn optimum combination of target angle θ_T and its preservation period for producing rapid position control towards assigned target position. The rapid position control is seen by the usage of target angle θ_T for producing acceleration and deceleration in achieving particular target position x_T. Therefore, it is understood that Learning Control System by multiple control knowledge was able to perform a rapid position control.

3.3 Application of Multiple Control Knowledge in Learning Control: Rapid Position and Obstacle Control

The second section of the Learning Control System by multiple control knowledge was designed for rapid position control with obstacle control of aerial hovering vehicles. Aerial hovering vehicles consist of non-linear parameters that require expertise in providing a safe and reliable control. Here, human expertise in operating such device is generated through application of Learning Control System by multiple control knowledge particularly in application of rapid position and obstacle control.

3.3.1 Parameters of Learning Control for Rapid Position and Obstacle Control

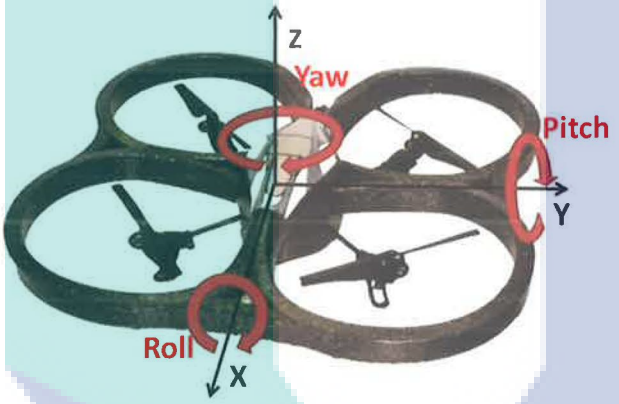


Figure 3.14: The angular dynamics of aerial hovering vehicle. (ArDrone by Parrot)

Continuing the Learning Control System design in chapter 3.2, three angle dynamics of the aerial hovering vehicle is concerned in designing the Learning Control System by multiple control function. In an unknown environment, it is difficult to perform a successful and optimum control operation due to availability of obstacles and other constraints. Here, Reinforcement Learning is applied to rewrite the control knowledge by determining the favourable state s; location and velocity, for an action a, which is the optimum target angular orientation θ_T for rapid position control while considering the existing obstacles. The control knowledge Q is updated using Q-learning as (3.3) and (3.4), which is

$$Q(s, \theta_T) = (1 - \alpha)Q(s, \theta_T) + \alpha[rew + \gamma Q_{max}], \tag{3.3}$$

$$Q_{max} = \max_{\theta'_T} Q(s', \theta'_T) \tag{3.4}$$

Where s and s' denotes state and future state of the control device, α is Learning Rate, γ is the discount rate and r is the reward.

However, as shown in figure 3.14, the aerial hovering vehicle applies three parameters of angular orientation, therefore, 3 optimum target angle must be learned in order to perform a rapid position control. Plus, effective combination of three target angles may help perform an optimum rapid position control around obstacles. In this case, target angle θ_T is a set of three target angles from the three parameters of angular orientation, as

$$\mathbf{\Theta_T} = \{\theta_{roll}, \theta_{pitch}, \theta_{yaw}\}.$$

From above, a set of 3 independent control knowledge Q is created for each target angle, as

$$\mathbf{Q} = \{Q_{roll}, Q_{pitch}, Q_{yaw}\}.$$

Since there are three sets of independent control knowledge will be used in the Learning Control System based on three dimensional angular orientation, state s were prepared to be three dimensional coordinates and velocities. State s consisted location $\mathbf{r}$, where

$$\mathbf{r} = \{x, y, z\},$$

and velocity according to each axis, $\mathbf{v}$, where

$$\mathbf{v} = \{v_x, v_y, v_z\}.$$

Therefore, state s is denoted as

$$s = \{\mathbf{r}, \mathbf{v}\}.$$

The reward rew used to update the control knowledge $\mathbf{Q}$ is based on (3.5),

$$rew = \frac{d_s - d_{s'} + 1}{d_{s'}} \tag{3.5}$$

where d_s is the distance between the control device at state s and the target location, and $d_{s'}$ is the distance between the control device at state s' and the target location, as shown in figure 3.15.

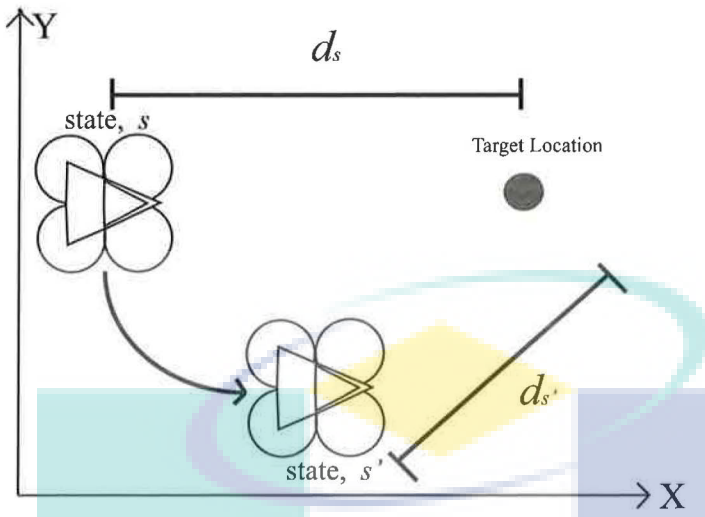


Figure 3.15: Parameters for determining rewards in Learning Control System for rapid position control.

Reward r in (3.5) was applied for two reasons; to have the control device travel a large distance between two states, and to have the control device distinguish the favourability of states that are closer to target position. This is because, larger travel distance between two states represent higher acceleration that was needed for performing rapid position control for reaching the target state at a faster rate.

Besides (3.5), reward *rew* is a constant in case of the Learning Control System failed to reach the target state within the designated simulation time, and when the control device exceed the designated movement range for the simulation.

3.3.2 System Structure for Rapid Position and Obstacle Control

Using the learning function arranged in the previous section, a Learning Control System by multiple control knowledge concerning application of three target angles was designed. The design of Learning Control System by multiple control knowledge for rapid position and obstacle control is as shown is figure 3.16, where three control knowledge for producing three target angles were applied. The design of the Learning Control System learns the optimum combination of target angles with predetermined preservation time of target angles and constant elevation. The Learning Control System was design to provide controls of position transition for 2 dimensional environment using application of three target angles assigned in the control knowledge of the Learning Control System.

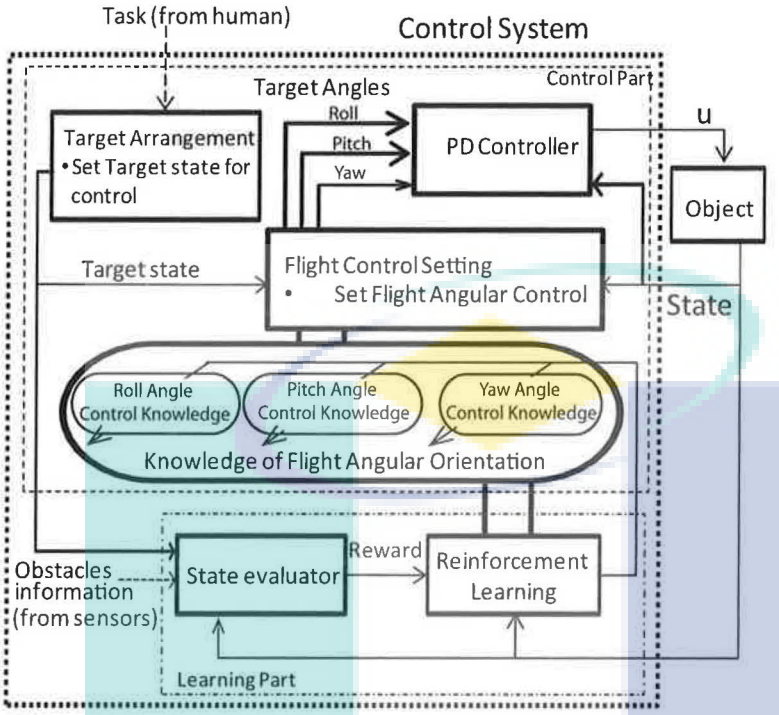


Figure 3.16: The structure of the Learning Control System for rapid position control.

3.3.3 Simulation Settings

Series of simulations were conducted in MATLAB Simulink based on the parameters of the aerial hovering vehicles shown in figure 3.14. These parameters are shown in table 3.4. A series of simulations which consisted different target position was assigned to confirm the effectiveness of the Learning Control System. Obstacles were also assigned in the simulation to confirm that the Learning Control System was able to operate through obstacles as intended. The assigned target states and obstacles were placed as shown in figure 3.17.

Table 3.4: Specifications of the simulated aerial hovering vehicle.

Parameters	Value
Weight	0.42 [kg]
Size:	
Length	0.53 [m]
Width	0.52 [m]
Height	0.1 [m]

The parameters for Q-learning is as shown in table 4.2, where these parameters were

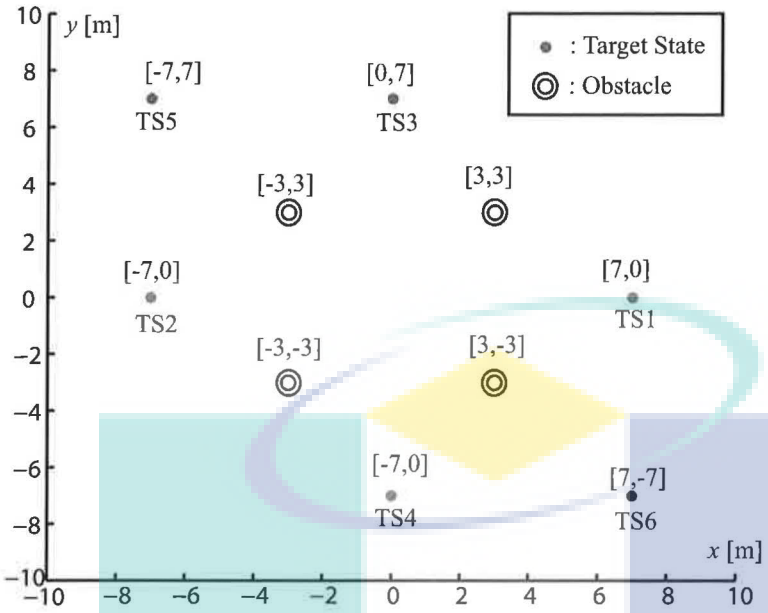


Figure 3.17: Obstacles and target location assigned in the simulations of Learning Control System for rapid position control.

selected pre-simulation. The position control of the aerial hovering vehicle was only applied on horizontal movements with constant altitude, within a movement range assigned.

Table 3.5: Q-learning parameters of Learning Control System for rapid position control.

	Parameters	Range	Intervals
State	Location, r[m]	$-10 < r(x,y) < 10$ $r(z) = 1$	2
	Velocity, v [m/s]	$-10 < v < 10$	2
Action	Target Angle, θ_T[rad]	$-0.25 < \theta_T < 0.25$	0.05

Learning rate, α	0.5	Discount rate, γ	0.3

There are several properties designated into the conducted simulations. For each simulation for each target state, the properties are as follows.

- Simulation runs six times with different target state assigned with each having 4 four permanent cylindrical obstacles with diameter of 1[m].
- Simulation end at 3000 episodes of trials.

- 30 second operation time for each episode.
- Action is evaluated for reward and target angles were renewed every 1 second.
- $\varepsilon-$greedy selection of each target angles
- $rew = -2$ when the action leads to out of range or obstacles.
- Due to large intervals on states, the controller for states within 1[m] around the target state will be switched to PD control.

The results from the simulations are determined by the accumulated rewards through the simulations and the successful attempts on achieving the target position by operating with and without obstacles.

3.3.4 Simulation Results

At the end of the simulation, the result of the trials for each episode was collected and analyses to confirm the reliability of the system. The results should provide the information on the control path for each target state assigned. This includes position transition and angular transition which is important for distinguish the reliability of the Learning Control System, with or without obstacles in the environment. The results also provide information regarding the improvement occurred in the control knowledge. Therefore, the results of the simulation are viewed in two aspects. The first aspect is the characteristic of rapid position control operation that successfully operates within an environment while the second aspect is the improvement of control knowledge that is used to perform the rapid position control.

3.3.4.1 Successful Control Operations towards Designated Target States

This result confirms the reliability of the Learning Control System for performing successful control operation that is required to reach the assigned target state. There are 6 target states were assigned with the same initial starting position in the simulation. A successful control attempt for each target states that was learned by the system during the simulation is shown in figure 3.18. Figure 3.18 shows the control operation that was accomplished at the final, 3000th episode of the simulation for each target position assigned.

The results show that the Learning Control System was able to control the control object towards each designated target states. Simulation for target 1 to 2 shows that direct movement from start position was able to achieved, when the movement path is not

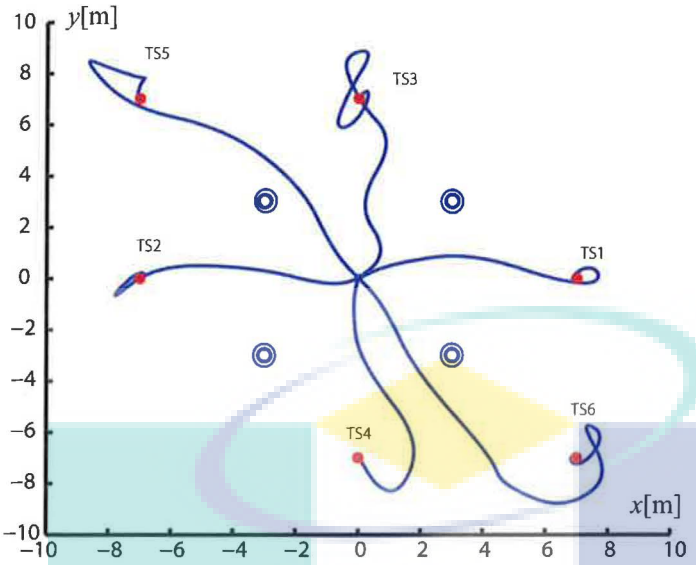
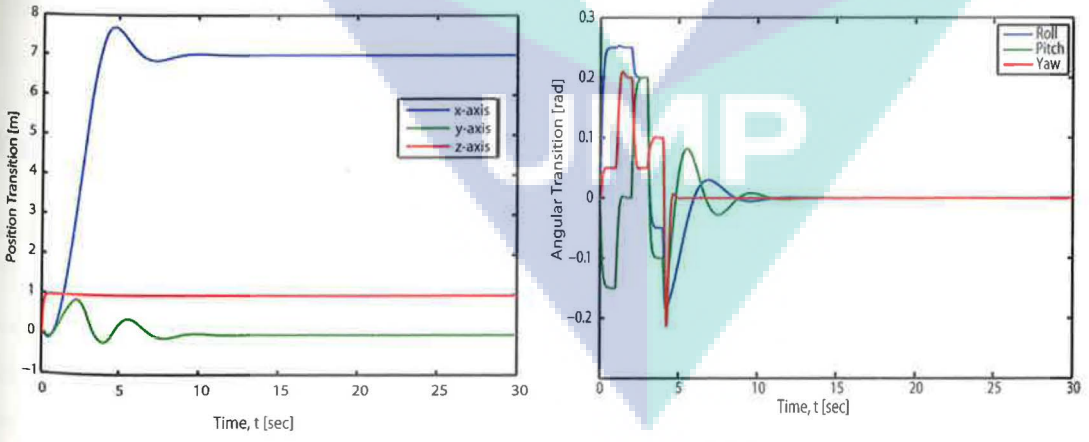


Figure 3.18: Successful control operation for the simulation with assigned target state.

obstructed by any obstacles. However, for target 3 and 4, the movement path was not so smooth compared to target 1 and 2. This is because, the system learns the most effective manoeuvres, and in case for target 3 and 4, the optimum manoeuvres that were learned here were not as smooth as for target 1 and 2, in figure 3.18. For target 5 and 6, the control system bent the movement path so that the control device can avoid the obstacles, but still reaches the assigned target state.



(a) Position transition.

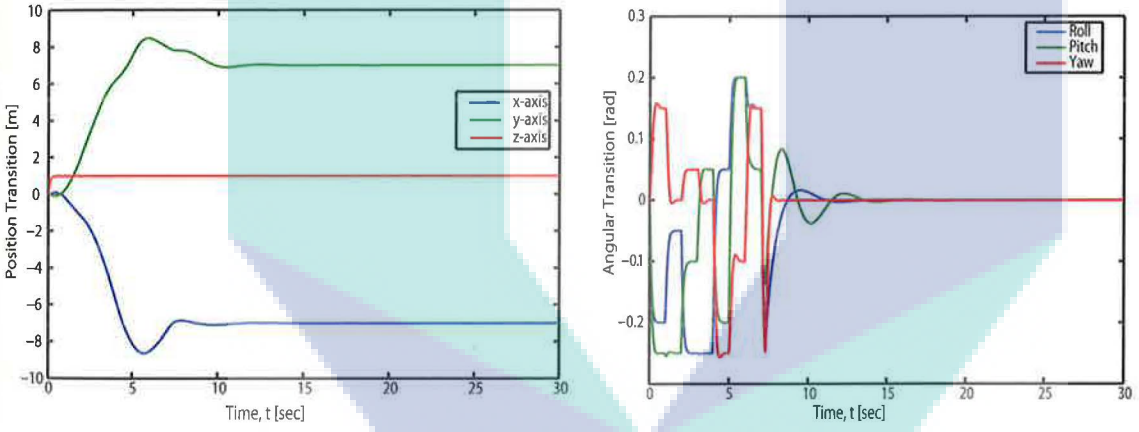
(b) Angles transition.

Figure 3.19: Successful control operation without obstacles in direct path. (Target State 1)

3.3.4.1.1 Successful Control Operation without Obstacles in Direct Path

This result explains the movement path of the control device that was operated by the system towards reaching target state 1. The direct path towards target state 1 is unblocked by any obstacles but the Learning Control System is needed to be careful of the obstacles at the side of the direct path. The details of the control operation for reaching target state 1 is shown in figure 3.19a and figure 3.19b.

Figure 3.19a shows the position transition of the control device in each 3 axis, during the final episode of simulation for Target State 1. Here, the system selects the optimum position transition for achieving the target state, with less unnecessary movements according to each axis. Figure 3.19b shows the transition of angular orientation based on roll pitch and yaw during the final episode of simulation for Target State 1. Here, the manipulation of angle can be seen to influence the position transition in figure 3.19a.



(a) Position transition.

(b) Angles transition.

Figure 3.20: Successful control operation with obstacles in direct path. (Target State 6)

3.3.4.1.2 Successful Control Operation with Obstacles in Direct Path

This result explains the movement path of the control device that was operated by the Learning Control System towards reaching target state 6. The direct path towards target state 6 is blocked by an obstacle and the system is needed to consider this obstacle when performing control operation to reach target state 6. The details of the control operation for reaching target state 1 is shown in figure 3.20a and figure 3.20b.

Figure 3.20a shows the position transition of the control device in each 3 axis, during

the final episode of simulation for Target State 6. Here, the system selects the optimum position transition for achieving the target state, with necessary movements according to each axis, needed to avoid the obstacles place in the environment. Figure 3.20b shows the transition of angular orientation based on roll pitch and yaw during the final episode of simulation for Target State 6. Here, the manipulation of angle can be seen to influence the position transition in figure 3.20a for taking necessary movements to avoid the assigned obstacle.

3.3.4.2 Control Knowledge Improvements during Control Operations towards Designated Target States

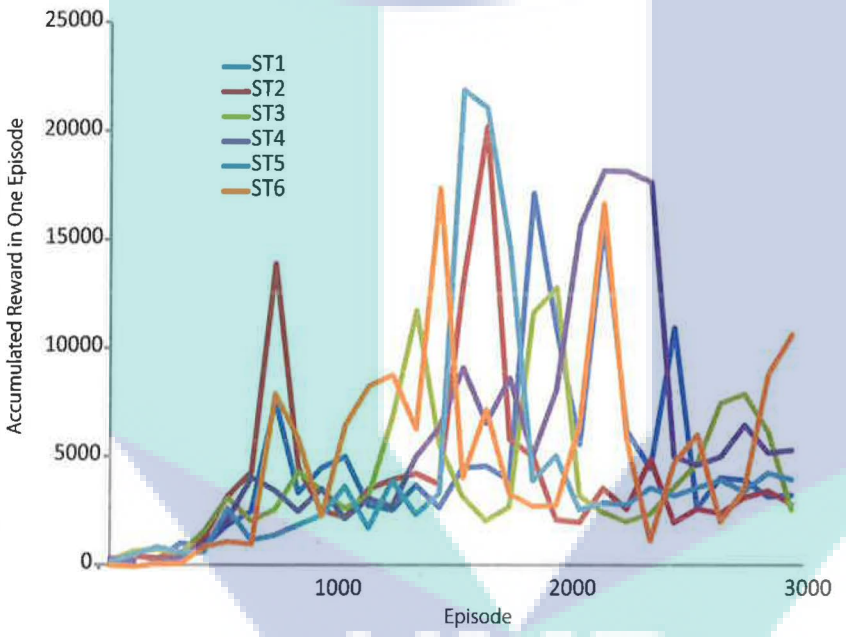


Figure 3.21: Accumulation of reward during simulations of Learning Control System for rapid position control.

This result shown in figure 3.21 explains the improvement that occurred during the simulation of the Learning Control System. For each episode, Control Knowledge has been updated to satisfy the environment where the control operation will be performed. Therefore, the increasing number of accumulated rewards represents increasing number of successful control operation. This explains that the Learning Control System learned the best control operation needed by attempting the control operation that leads to most reward in each episode, where successful control attempts were learned during the simulation

that leads to more reward accumulated through more episodes. Here, multi-functionality was achieved by manipulation of aerial hovering vehicles in rapid position control around obstacles.

3.4 Summary

Human operation of Rapid Position Control applies multi-functionality to control an Aerial hovering vehicle with precision and safety. Multi-functionality applies multiple knowledge of control in providing such precision and safety in controlling devices especially devices with non-linearity. Providing Learning Control System with multiple control knowledge may provide multi-functionality in controlling complex non-linear device where human can expertly controls due to multi-functionality in human control ability.

In this chapter, Learning Control System by Multiple Control Knowledge was designed and applied on rapid position control of aerial hovering vehicle. The Learning Control System with Multiple Control Knowledge is designed for performing an operation that requires multiple functions in controlling a device. The Learning Control System was firstly designed and applied for rapid position control alone, using cart-pendulum system as control device. It was later designed for control of aerial hovering vehicles for rapid position control among vehicles.

Simulations were conduct to confirm application of multi-functionality in rapid position control using the designed Learning Control System on aerial hovering vehicles. The controls of aerial hovering vehicles were emulated on cart-pendulum system, where Learning Control System for rapid position control was designed, before being applied on simulation of the aerial hovering vehicle. Simulations show that the control object has multiple control functions to learn and to control for performing rapid position control while considering surrounding obstacle to reach the assigned target state. Development of a Learning Control System with multiple sources of control knowledge provides multi-functionality in rapid position control while considering obstacles on an aerial hovering vehicle. Therefore, the design of Learning Control System with multiple control knowledge may provide multi-functionality in controls application.

Chapter 4

Multiple-Functions Learning Control by Compound Function

Design of Learning Control with quality of multi-functionality produces functions that require decision management in order to optimize the usage of each function. Providing decision management requires a Learning Control System to be able to analyse surrounding environment and considers necessary function required by particular specifications of the environment. Compound Function provides multi-functionality with decision management, where necessary function is provided based on the environment that requires them. Design of Learning Control System with Compound Function may render a control device autonomous in control operation due to decision management properties that provide action consideration during the operation.

4.1 Compound Function

A human has the ability to learn and utilize their skills from experiences when confronting any problem. Such ability capable those in utilizing certain knowledge of skills that was obtained through various experiences for solving a new problem that requires a configuration of obtained skills. In case of hurdle race, human can utilize the skills of jumping and running into performing hurdle race. Both control knowledge of jumping and running must be utilized by optimum configuration in order to provide an effective hurdle operation. Here, human utilizes this knowledge of skills in creating an action by considering the requirement of each skill, as shown in figure 4.1. The above skills are not only being utilized in solving problems, however, the executed actions may provide feedback and help develop the skills that were performed through development of control knowledge of those skills.

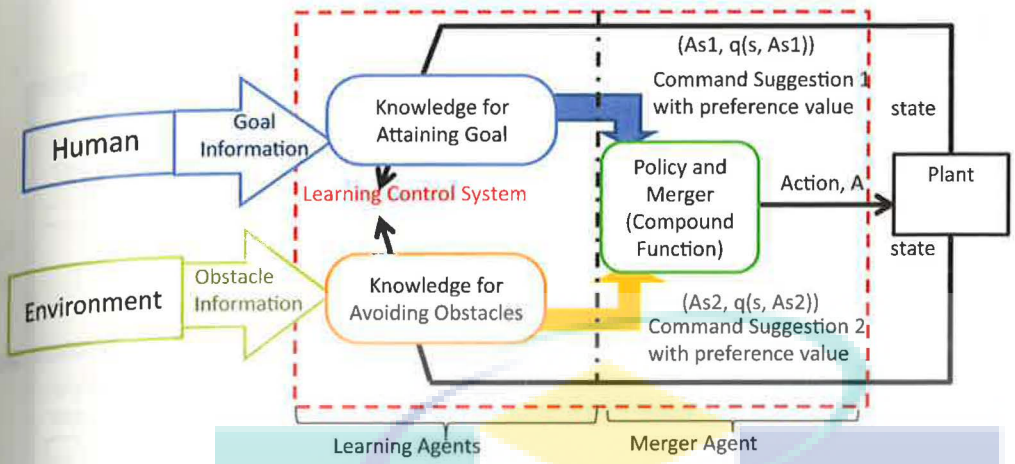


Figure 4.1: System structure for application of compound function.

Based on figure 4.1, there are two agents involves in creating a Learning Control System with compound function. The first agent is the learning agent, where control functions where arranged in the system. The second agent is the merger agents, where compound function merges the control information provided by the control functions within the learning agent. Compound Function is created in order to provide such human ability in a control system. Learning Control System may provide updates to control knowledge however, when having multiple control knowledge in a Learning Control System, consideration of control functions is needed to determine the control knowledge that provides this function. Compound Function provides consideration in selecting the best control knowledge for applying necessary control function through the Learning Control System. A suggestion of control command together with the preference value is provided by the two control knowledge, where the compound function considers the optimum action for operating the control device. The feedback of the action will provide update for the control knowledge of the operated action, enhancing them for consideration in future operation. Here, a design of Learning Control that utilizes multiple functions by Compound Function was utilized for obstacle and goal consideration of a mobile robot. The Compound Function merges the control knowledge from each control function and stores the control information obtained from the source control knowledge for evaluation in form of compound control knowledge. The Learning Control System was designed to apply the proposed Compound Function to determine the priority of the control source in executing action based on two Control Knowledge of Goal Attainment and Obstacle Avoidance.

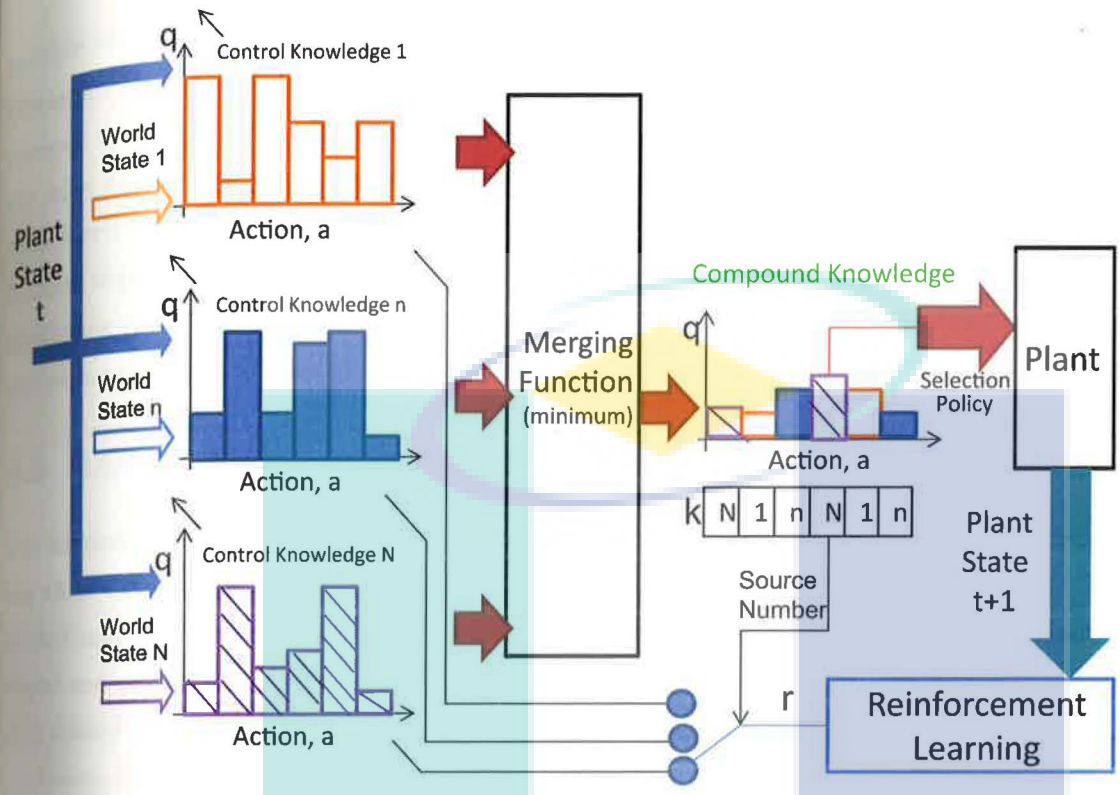


Figure 4.2: Method of multiple functions Learning Control by compound control knowledge.

4.2 Compound Control Knowledge

In order to create a Learning Control System that can utilize multiple Control Functions, Control Knowledge from each Control Functions must be merged into one Compound Control Knowledge. The Compound Control Knowledge proposed in Fig. 4.2 can be applied to two or more Control Function. However, in order two confirm the validity of the compound control function, two control functions was applied on the Learning Control System.

The compound control knowledge was created through selecting the minimum option of action compared based on the preference value provided by control knowledge of each control function. A new set of action is obtained, consists the minimum preference value obtained from comparing both control knowledge. Action of the control device is selected through the compound control knowledge where the action with optimum value among the action with minimum value stored in the compound knowledge is selected. Updates are return to the control knowledge of control function that provided the executed action.

Hierarchical Reinforcement Learning applies comparison between preference values for

multiple value functions similar to compound function [47]. However, the application of Hierarchical Reinforcement Learning requires layers of value functions where each layer is surveyed by parent task. Application of value function in the lower layer is determined by the value function in the upper layer of the hierarchy [42]. In case of compound function, value functions are arranged without hierarchy, where the application of the necessary value function depends on the state of the control device. The value function is merged when application of more than one value functions is necessary through merging function where minimum preference value between the value functions is selected.

4.3 Learning Agent for Compound Function Device

The learning process applied in the Learning Control System consists Reinforcement Learning where Control Knowledge is updated in a form of value functions Q. The value function of the Control Knowledge is denoted by state S, defining the current situation of the control object and action A, defining the following move of the control device. State S and Action A is defined into two sets as State $S = \{s_1, s_2, .., s_n\}$ and Action $A = \{a_1, a_2, .., a_n\}$.

During the phase of updating the control knowledge, the preference value q of the combination between state s and action a is renewed by the reward r obtained after performing the action a. In the case of successful operation, the preference value q increases, and decreases in result of unsuccessful operation. The value function of the Control Knowledge is updated based on Q-learning algorithm shown in equation 1.1. Here, two Learning Agents for Compound Function Device was designed using Reinforcement Learning; the first Learning Agent consists Learning Control System for goal attainment function, while the second Learning Agent consists Learning Control System for obstacles avoidance.

4.3.1 Learning Control System for Goal Attainment Function

Learning Control System for goal attainment operates the control device towards the goal. Here, the Learning Control for goal attainment applies goal distance $\Delta G = \{\Delta X_G, \Delta Y_G\}$ as state S while movement distant $\Delta\tau$ and rotation θ as action A_G. Therefore, the value function Q for Learning Control for Goal Attainment is defined by $Q(\Delta G, A_G)$.

The update equation for the Learning Control System for goal attainment alone is,

$$Q_1(\Delta G, A_G) = (1 - \alpha)Q_1(\Delta G, A_G) + \alpha[r + \gamma_1 Q_{max}], \tag{4.1}$$

$$Q_{max} = \max_{A_G} Q_1(\Delta G, A_G) \tag{4.2}$$

where reward r is assigned according to the function shows in figure ??. Here, rewards are given according to distance between the goal and the control device. Action that renders the device further than goal will result in negative rewards while action that renders the device closer will result in positive reward.

4.3.2 Learning Control System for Obstacle Avoidance Function

Learning Control System for obstacle avoidance operates the control device away from obstacles. Here, the Learning Control System for obstacle avoidance utilizes obstacle distance $\Delta O = \{\Delta X_O, \Delta Y_O\}$ as state S and movement distant τ and rotation θ as action A_O. Therefore, the value function Q for Learning Control System for obstacle avoidance is defined by $Q(\Delta O, A_O)$.

The update equation for Learning Control System for obstacle avoidance alone is,

$$Q_2(\Delta O, A_O) = (1 - \alpha)Q_2(\Delta O, A_O) + \alpha[r + \gamma_2 Q_{max}], \tag{4.3}$$

$$Q_{max} = \max_{A_O} Q_2(\Delta O, A_O) \tag{4.4}$$

where reward r is assigned according to the function shows in figure ??. Here, rewards are given according to distance between the obstacle and the control device. Action that renders the device further than detected obstacle will result in positive rewards while action that renders the device closer will result in negative reward.

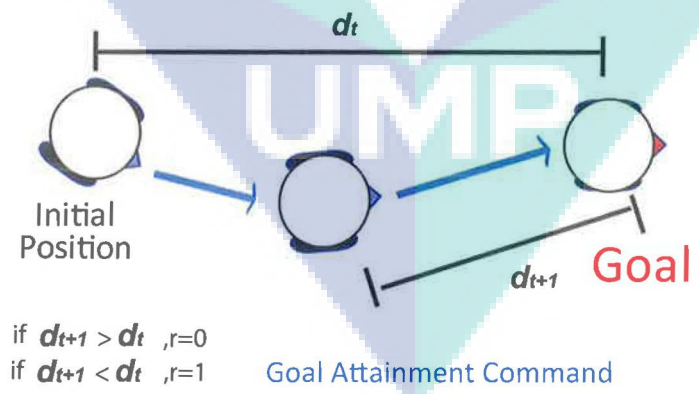


Figure 4.3: Reward for control in Goal Attainment Function.

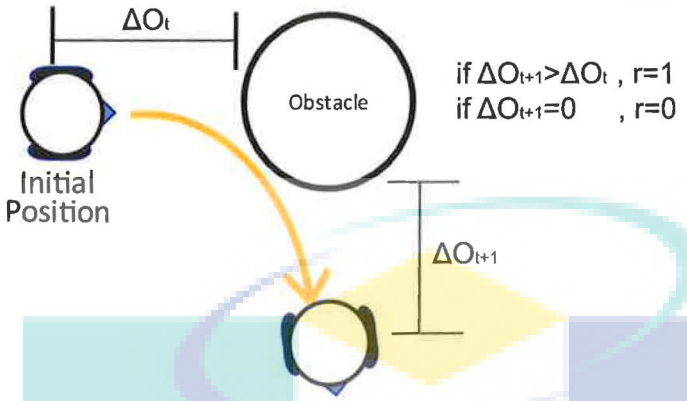


Figure 4.4: Reward for control in Obstacle Avoidance Function.

4.4 Merger Agent for Compound Function Device

Having two or more Control Functions in one Learning Control System would require the system to utilize the value function from both Control Functions. The preference value from both value functions is used to describe the priority in selecting the best actions provided by each value functions. In order to provide comparisons between two or more value functions, the update method for the participating value functions has a limit between 0(bad) and 1(Good). Therefore, the discount rate γ of the updated value in equation 4.1 and 4.3 for each value functions is applied as,

$$\gamma_1 = 1 - Q_1(\Delta G, A_G), \tag{4.5}$$

$$\gamma_2 = 1 - Q_2(\Delta O, A_O). \tag{4.6}$$

A new value function defines as Compound Control Knowledge is firstly constructed using the value functions provided by the Learning Agent as shown in figure 4.5. The value function of Compound Control Knowledge is constructed by Q_{All} and K,

$$Q_{All} = \underset{n=1,2}{Min} Q_n(S_t, A) \tag{4.7}$$

with

$$K(s_t, A) = n, \tag{4.8}$$

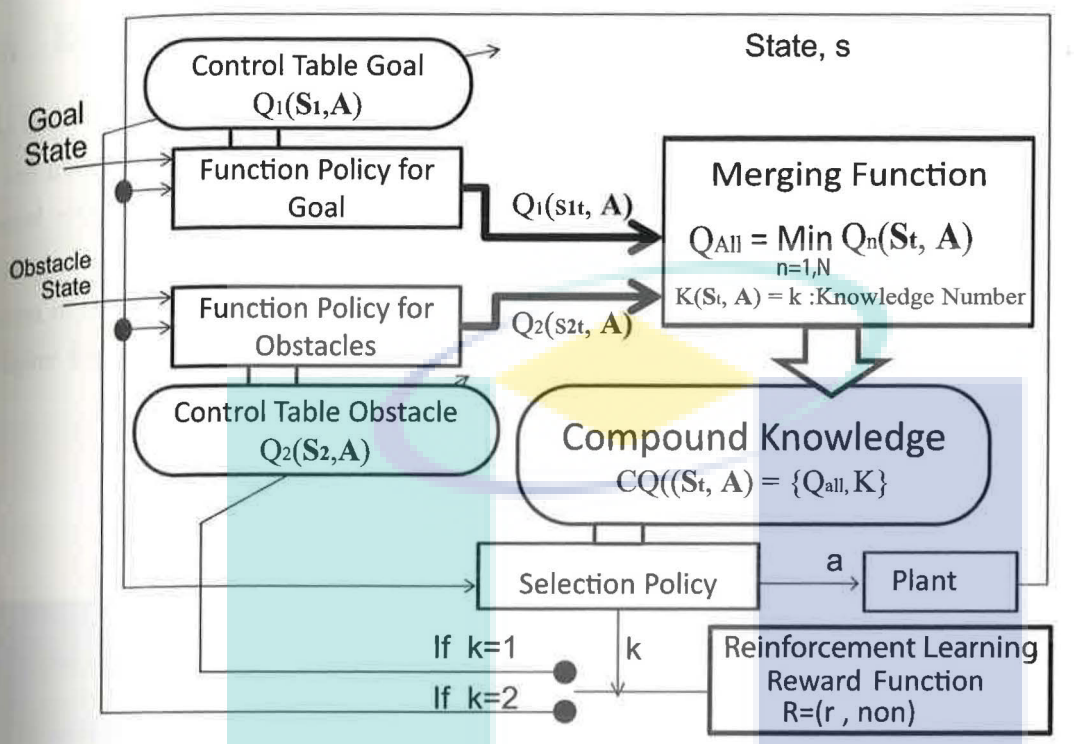


Figure 4.5: Structure of Learning Control System by compound function in case of 2 control functions. (Goal and Obstacle)

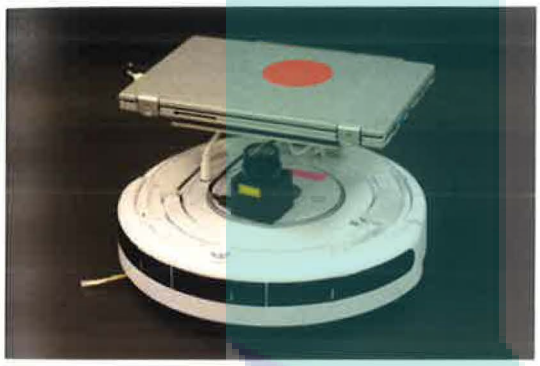
where n is the serial number of the source value functions in the subsystem, which defines the Compound Control Knowledge as

$$CQ(S_t, A) = \{Q_{All}, K\}. \tag{4.9}$$

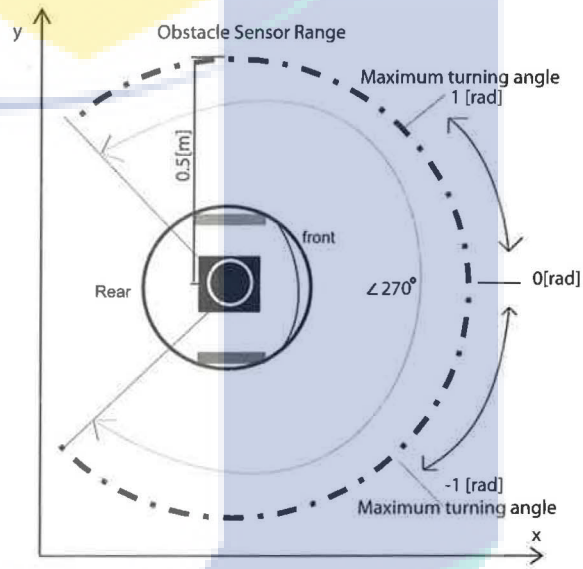
Based on the above equations, the overall design of Learning Control System by Compound Function for goal attainment and obstacle avoidance is as shown in figure 4.5. Here, Learning Agents supplied control information into the merging function, where a new value function defined as compound control knowledge is created. Action is selected through the compound control knowledge and the Reinforcement Learning Function updates the Learning Agents depending on the source of the executed action. The effectiveness of the designed Learning Control System by Compound Function was confirmed through series of experiments, where the design was applied on control operation of a small mobile robot among obstacles.

4.5 Experiments Settings

The Learning Control System by Compound Function shown in figure 4.5 was evaluated in two phases; simulation phase and experiment phase. During the simulation phase, the control object applied was a robot that was designed based on parameters as the robot shown in figure 4.6a. The robot shown in figure 4.6a was applied for evaluating the designed system in the experimental phase. The operation specifications of the robot are as shown in figure 4.6b, while the physical specifications of the robot are as shown in table 4.1.



(a) Robot for real operation experiment.



(b) Robot structure for simulation and real operation experiment.

Figure 4.6: Specification of the control device for experiments.

Table 4.1: Specifications of the simulated control device for Learning Control System by compound function.

Parameters	Value
Weight	5.5 [kg]
Size:	
Length	0.27 [m]
Width	0.27 [m]
Height	0.15 [m]

The simulation was conducted as a platform to train the control knowledge of the Learning Agents in the Learning Control System and for the evaluation of compound function

application. The simulation environments are based on the field map in figure 4.7. The parameters for the equation in the Learning Control System are as shown in table 4.2. The simulation was conducted in three phases; two phases for training and one phase for evaluation. The phases of training were conducted each to construct the Control Knowledge for control functions of Goal Attainment and Obstacle Avoidance. The phase of evaluation was conducted in evaluating the effectiveness of the compound function in applying the two Learning Agents. The training phases were conducted in 750 episodes, with 5 targets, while the evaluation episode was conducted in 375 episodes for 5 goals. The results obtained concerning the movements of the robot and the condition of the learning process was evaluated.

Table 4.2: Parameters for Q-learning in Learning Control System by compound function.

	Parameters	Range	Intervals
State (Goal)	Goal Distance,ΔG[m]	$-10 < \Delta G(x, y) < 10$	2
State (Obstacle)	Obstacle Distance,ΔO[m]	$-2 < \Delta O(x, y) < 2$	0.5
Action	Target Angle, θ[rad] Travel Distance, V[m]	$-1 < \theta < 1$ $0.1 < x < 0.5$	0.5 0.2

Learning rate, α	0.5	Discount rate, γ	0.3

4.6 Experiments Results

The training phase describes the effectiveness of the control knowledge applied in the Learning Agents. The evaluation phase describes the effectiveness of the compound function utilizing the whole Learning Control System. The successful simulation obtained during the evaluation phase was applied on the robot. The robot movement was recorded and the effectiveness of the system in a real environment was evaluated as well.

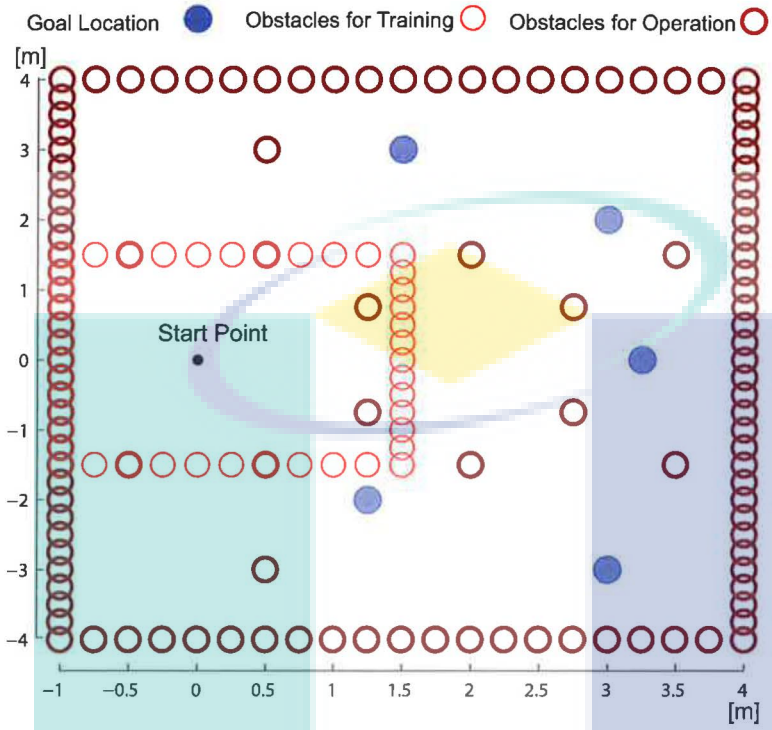


Figure 4.7: Field map for simulation of Learning Control System by compound function.

4.6.1 Simulation Results for Goal Training

Here, results based on the training of control knowledge for goal attainment provides information regarding the effectiveness of the Learning Control System in creating the control knowledge for obtaining goals. The control knowledge for goal attainment is important to provide comparison when applying the compound control knowledge.

Figure 4.8 and figure 4.9 describes the results of the training process for the Control Knowledge of Goal Attainment in the Learning Agent. In figure 4.8, the robot in the simulation was able to reach the target assigned. Movement strategies were constructed depending on the direction of the targets under the restriction of the assigned control command. Figure 4.9 shows the accumulated reward by the value functions of the control knowledge. The accumulated reward increases over episodes, where successful attempts towards the goals are achieved more frequently after several trials for each assigned target. Therefore, it can be concluded that the Control Knowledge of the Learning Agent for the Goal Attainment was successfully constructed.

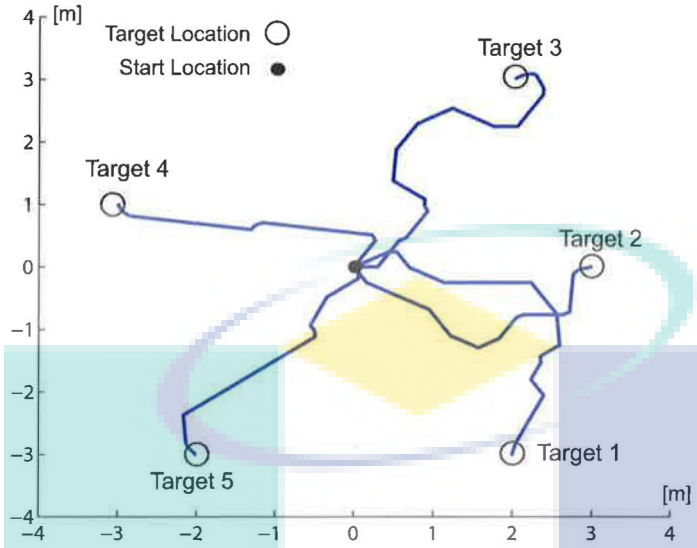


Figure 4.8: Training operation for achieving goal using Learning Control.

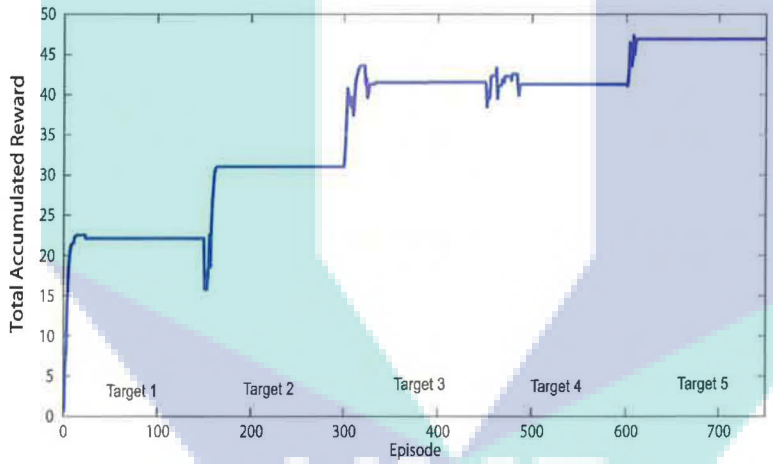


Figure 4.9: Accumulated reward for goal knowledge over simulation episode.

4.6.2 Simulation Results for Obstacles Training

Here, results based on the training of control knowledge for obstacles avoidance provides information regarding the effectiveness of the Learning Control System in creating the control knowledge for avoiding obstacles. The control knowledge for obstacle avoidance is important to provide safe control when applying the compound control knowledge.

Figure 4.10 and figure 4.11 describes the results of the training process concerning the Control Knowledge of Obstacle Avoidance in the Learning Agent. In figure 4.10, the

robot movement was obstructed by obstacles from reaching the assigned target. Frequent obstruction has created alternative movement strategy for the robot to avoid the obstacles as long as possible. Therefore, a successful Control Knowledge for avoiding an obstacle was obtained at the end of the simulation. Here, figure 4.11 shows that the accumulated rewards increases in the value function of the Control Knowledge of the Learning Agent for Obstacle Avoidance.

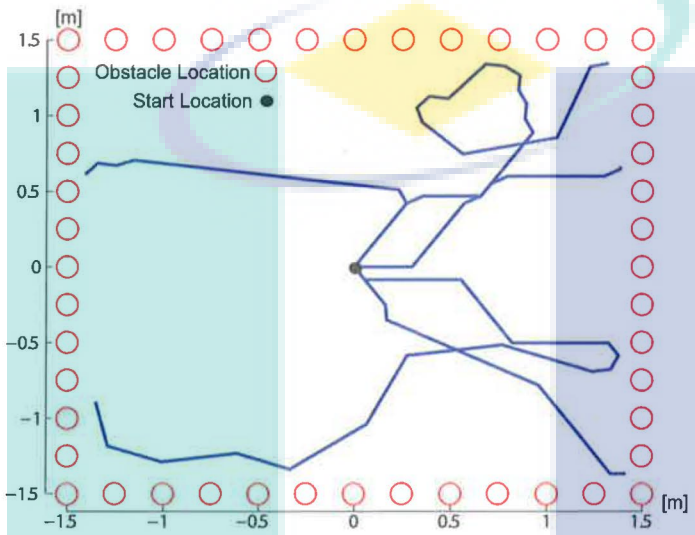


Figure 4.10: Training operation for avoiding obstacles using Learning Control.

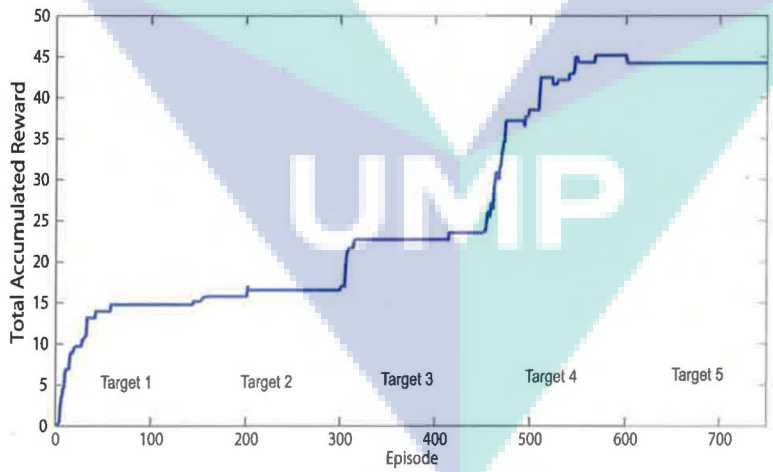
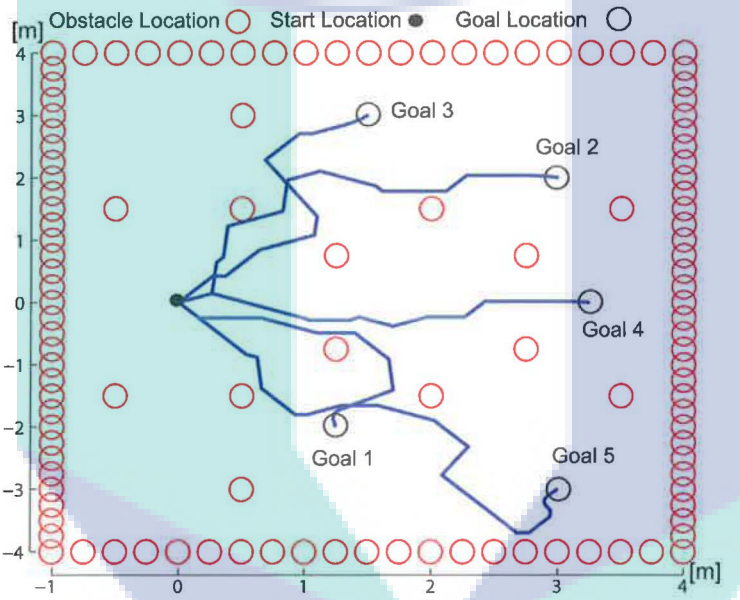


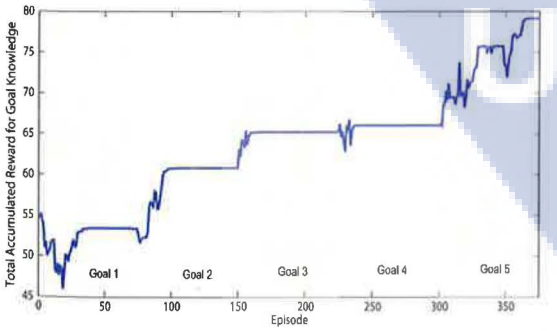
Figure 4.11: Accumulated reward for obstacle knowledge over simulation episode.

4.6.3 Simulation Results for Compound Function Training

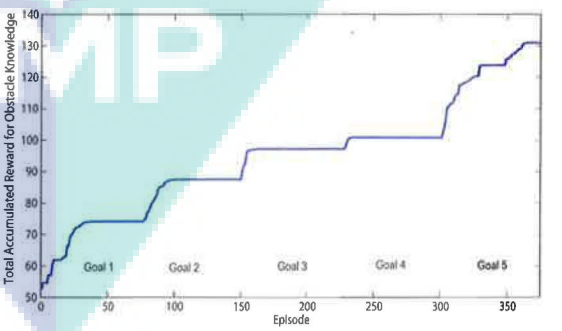
Figure 4.12 describes the results for the evaluation phase with obstacles and target. This result evaluates the effectiveness in creating the Compound Knowledge. Figure 4.12a shows the robot movements in the simulation where the robot was able to successfully reach all the assigned goals while avoiding all the obstacles. Figure 4.12b and Figure 4.12c described the changes in the Control Knowledge of Goal Attainment and Obstacle Avoidance in the Learning Agents. The value function of each Control Knowledge improves over the episodes. Therefore, the proposed system was effective in utilizing Learning Agents in performing a control operation for attaining goal while avoiding obstacles.



(a) Training operation for compound knowledge.



(b) Accumulated reward for goal knowledge over simulation episode.

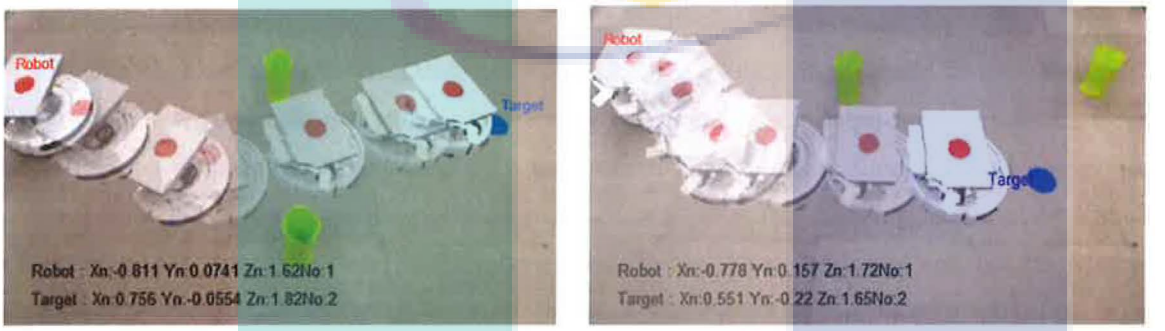


(c) Accumulated reward for obstacle knowledge over simulation episode.

Figure 4.12: Training results of compound knowledge in simulation.

4.6.4 Experiment Results on Real Operation

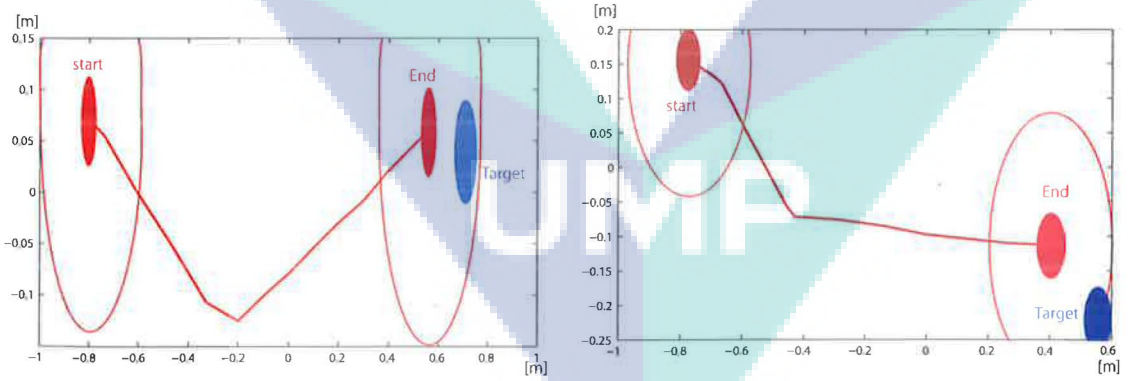
Figure 4.13 shows the results of the control operation in a real environment. The operation was conducted in a map monitored where the location data collected using Kinect for Windows. The obstacles and the target were assigned randomly and the movement of the robot was recorded. Figure 4.14 shows the movement configuration of the robot of figure 4.13. These results show that the robot successfully approaches the target position. The results confirm that the Learning Control System by Compound Function was effective in applying multi-functionality in goal attainment and obstacle avoidance on a control device.



(a) Operation with random obstacle and target.(case 1)

(b) Operation with random obstacle and target.(case 2)

Figure 4.13: Evaluation of real operation with robot.



(a) Operation with random obstacle and target.(case 1)

(b) Operation with random obstacle and target.(case 2)

Figure 4.14: Movement results of the evaluation.

4.7 Summary

Learning Control System with multiple functions requires decision management in order to provide multi-functionality in control operation. Applying decision management in Learning Control System may provide consideration in applying a control function among the option of control functions. Due to application of the decision management, necessary control function can be provided depending on the environment situation, increasing the reliability of control operation in any environment. Therefore, Learning Control System with multiple control functions requires a method for decision management in order to provide multi-functionality effectively.

In this chapter, a multi- function Learning Control System is designed to provide multi-functionality with decision management through application of Compound Function. Compound Function described as Merging Agent; consisting merging function and compound control knowledge may provide decision management through merging the control knowledge of control functions, described by Learning Agents, into compound control knowledge. Compound control knowledge is created through selecting the minimum preference value of action options when comparing Learning Agents. Application of compound function created new temporary compound control knowledge using elements from multiple control knowledge of control functions.

Series of experiment was conducted in order to confirm the effectiveness of the designed system. Two phases of simulation were conducted to construct the Learning Agents and to evaluate the Merging Agent. Results show that construction of Learning Agent was successful and was applied in the simulation for evaluating Merging Agent. Results of the evaluation phase show that the designed system was able to utilize compound function into applying multi-functionality during control operation for goal attainment and obstacle avoidance. Simulation results show that the system was able to apply the compound function in providing multi-functionality in form of Goal and Obstacle Consideration. Therefore, a Learning Control System with multiple functions was obtained with application of the Compound Function in the Learning Control System.

Chapter 5

Conclusion

Human actions involve multi-functionality, where an action could provide results for multiple purposes. Through this quality, consideration on multiple parameters can be made before an action can be executed. Providing such quality to a control system would require application of multiple control function under one system. A control system that is adaptable to environment with multi-functionality would render the control device autonomous in performing control operation. Therefore, adaptable control system with multi-functionality may provide safer and reliable control for a control device in any environment. Designs concerning application of Learning Control System with multi-functionality are provided through this dissertation. Here, the design involves methods of applying Learning Control that provides multiple control function for providing safer and reliable control for control operation.

In chapter 2, the first design of multiple functions Learning Control System utilizes substitute target in providing control solution in a constrained non-linear device. Constrained Non-linear Learning Control system by substitute targets provides control solution to multi dimensional states in Non-linear Control device under constraints. Results show that the Learning Control System by substitute target was able to provide multi-functionality in a constrained non-linear control device through application on cart-pendulum swing up control among constraints. Therefore, multi-functionality was applied on non-linear control device by substitute target and a safe and reliable control was obtainable through multi-function learning control.

In chapter 3, the second design of multiple functions Learning Control System utilizes multiple control knowledge in providing control solution in controls of a non-linear device, consisting cart-pendulum system and aerial hovering vehicles. Learning Control System

by multiple control knowledge provides solution in applying human like control decision to a machine that reduces dependency in detailed human command. Results show that the Learning Control System by multiple control knowledge was able to provide human like multi-functionality in controls of non-linear device through application of multiple control knowledge in rapid position control of aerial hovering vehicles. As a result, the nonlinear control by the integration of multiple control knowledge in the learning control system were obtained, operated similar to human skills, thus the multivariable multi-function control was achieved.

In chapter 4, the third design of multiple functions Learning Control System utilizes compound function in providing decision management in Learning Control System with multiple control knowledge of functions. Compound Knowledge Learning Control system provides control solution for having control functions priority consideration in environment with multiple control functions. Results show that the Learning Control System by compound function was able to provide necessary consideration between application of goal attainment control or obstacle avoidance control during operation of a small robot device. Therefore, the compound knowledge (state-action rule) that integrates goal attainment function and the obstacle avoidance function was learned for providing multi-functional control.

In this research, Learning Control System with multi-functionality is designed and developed. By the designs, Learning Control System with multi-functionality may provide human-like safe and reliable control in a control device, making it capable of providing autonomous control in any environment.

Acknowledgements

The author would like to express his deepest acknowledgment to advisor, Professor Seiji Yasunobu for his constant guidance throughout his study. The acknowledgment continues to Dr. Takeshi Shibuya for advise and guidance on Reinforcement Learning and experimental analysis. Special appreciation goes to Professor Noriyuki Hori and Professor Hiromi Mochiyama for advises during research. Appreciation also goes to all members of the Intelligent Control System Laboratory for all the support and help given through out the research. Finally, his heartfelt appreciation to the Malaysian Government for the degree scholarship and also to all his family members for their constant support and encouragement.

Bibliography

[1] Tom M. Mitchell. "Machine Learning." *Portland, Oregon: McGraw Hill*, 1997.

[2] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell. "Machine Learning: An Artificial Intelligent Approach." *Los Altos, California: Morgan Kaufman Inc.*,1986.

[3] Ethem Alpaydin. "Introduction to Machine Learning." *Cambridge, Massachusetts: The MIT Press*, 2010.

[4] R. S. Sutton and A. G. Barto. "Reinforcement Learning: An Introduction." *Cambridge, Massachusetts: The MIT Press*, 1998.

[5] Thomas G. Dietterich. "Machine Learning." *Nature Encyclopedia of Cognitive Science*, London: Macmillan, 2003.

[6] Stefan Schaal and Christopher G. Atkeson. "Learning Control in Robotics." *Robotics & Automation Magazine*, Vol.7 Issue 2 pp.20-29, 2010.

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

[8] H. Yamasaki and S. Yasunobu. "Evolutionary Control Method and Swing Up and Stabilization Control of Inverted Pendulum." *IFSA World Congress*, pp. 2078-2083, 2001

[9] T. Matsubara and S. Yasunobu. "An Intelligent Control Based on Fuzzy Target and Its Application to Car Like Vehicle." *SICE Annual Conference*, pp. 2489-2494, 2005

[10] V. N. Vichugov, G. P. Tsapko and S. G. Tsapko. "Application of Reinforcement Learning in Control System Development." *8th Russian-Korean International Symposium on Science and Technology*, pp. 732-733, 2005

[11] N. Kazuhiro, T. Tsubone and Y. Wada. "Possibility of Reinforcement Learning Using Event-Related Potential Toward and Adaptive BCI." *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1720-1725, 2009

[12] S. Nakamura and S. Hashimoto. "Hybrid Learning Strategy to Solve Pendulum Swing-Up Problem for Real Hardware." *IEEE International Conference on Robotics and Biomimetics*, pp. 1972-1977, 2007

[13] M. Riedmiller. "Neural Reinforcement Learning to Swing-Up and Balance a Real Pole." *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 3191-3196, 2005

[14] D. L. Abel. "Constraints vs Controls." *The Open Cybernetics & Systemics Journal*, Vol. 4, pp. 14-27, 2010

[15] 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

[16] J. Valasek, J. Doebbler M.D. Tandale and A.J. Meade "Improved Adaptive-Reinforcement Learning Control for Morphing Unmanned Air Vehicles." *IEEE SMC*, pp.1014-1020,2008

[17] Y. Nakamura, S. Ohnishi, K. Ohkura and Kanji Ueda "Instance-Based Reinforcement Learning for Robot Path Finding in Continuous Space." *IEEE SMC*, pp.1229-1234,1997

[18] S. Schaal and C.G. Atkeson " Learning Control in Robotics; Trajectory-Based Optimal Control Techniques." *IEEE Robotics and Automation Magazine*, Volume 7 Issue 2 pp.20-29,2010

[19] X. Xu, L. Zuo and Z. Huang "Reinforcement learning algorithms with function approximation: Recent advances and applications." *35th Journal of. Information Sciences* Vol. 261, pp. 1-31, 2014

[20] D. Chang and J. E Meng "Real-Time Dynamic Fuzzy Q-Learning and Control of Mobile Robots." *35th 5th Asian Control Conference*, Vol.3, pp.1568 - 1576, 2004

[21] L. Busoniu, R. Babuska and B.D. Schutter "A Comprehensive Survey of Multiagent Reinforcement Learning." *IEEE Trans. Systems, Man and Cybernetics*, Vol.38 No.2, pp.156 - 172, 2008

[22] V. Gullapalli " Direct Associative Reinforcement Learning Methods for Dynamic Systems Control." *Neurocomputing 9*, pp.271 - 292, 1995

[23] R. Sun and C. Sessions "Multi-agent reinforcement learning with bidding for automatic segmentation of action sequences." *4th International Conference on MultiAgent Systems*, pp.445 - 446 , 2000

[24] J. Yu "An Adaptive Gain Parameters Algorithm for Path Planning Based on Reinforcement Learning." *4th International Conference on Machine Learning and Cybernetics*, pp.3557 - 3562, 2005

[25] S. G. Khan, G. Herrmann, F. L. Lewis, T. Pipe, and C. Melhuish " Reinforcement learning and optimal adaptive control: An overview and implementation examples." *Annual Reviews in Control*, 36 (1), pp. 42-59, 2012

[26] C. Obayashi, T. Tamei, and T. Shibata "Assist-as-needed robotic trainer based on reinforcement learning and its application to dart-throwing." *Neural Networks*, pp.52-60, 2014

[27] L. Panait, and S. Luke "Cooperative Multi-Agent Learning: The State of the Art." *Autonomous Agents and Multi-Agent Systems*, Vol.11, pp. 387434, 2005

[28] J. Choi, S. Oh, and R. Horowitz "Distributed learning and cooperative control for multi-agent system." *Automatica 45*, pp.2802-2814, 2009

[29] A. Kabysh , V. Golovko, and A. Lipnickas "Influence Learning for Multi-Agent System based on Reinforcement Learning." *Computing*, Vol. 11, Issue 1, pp.39-44, 2012

[30] Umano M., Ise A. and Seta K. "Tuning of Fuzzy Rules with a Real-Coded Genetic Algorithm in Car Racing Game." *27th Fuzzy System Symposium*, TC3-3, pp. 1-6, 2011.

[31] W.R. Ferrell and T.B. Sheridan "Supervisory control of remote manipulation." *Spectrum, IEEE* , Vol.4 , Issue 10, pp.81-88, 1967

[32] J. Rasmussen, "Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models." *IEEE Transactions on Systems, Man and Cybernetics*, Vol. SMC-13, pp. 257-266, 1983

[33] Christos G. Cassandras "Sensor Networks and Cooperative Control." *European Journal of Control*, Vol. 11, Issues 45, pp.436463, 2005

[34] M Burger and M.Guay "Robust Constraint Satisfaction for Continuous-Time Nonlinear Systems in Strict Feedback Form." *IEEE Transactions on Automatic Control* , Vol.55 , Issue 11, pp.2597 - 2601, 2010

[35] K. Doya "Reinforcement Learning in Continuous Time and Space." *Neural Computation 12*, pp.219245 , 2000

[36] L. P. Kaelbling, M. L. Littman and A. W. Moore "Reinforcement Learning: A Survey." *Journal of Artificial Intelligence Research*, Vol. 4, pp.237-285, 1996

[37] C. Watkins and P. Dayan "Technical Note Q-Learning" *Machine Learning*, Vol. 8, pp.279-292, 1992

[38] K. Nomoto, T. Tsubone and Y. Wada "Possibility of reinforcement learning using event-related potential toward an adaptive BCI." *IEEE International Conference on Systems, Man and Cybernetics*, pp.1720-1725, 2009

[39] M. Riedmiller, "Neural reinforcement learning to swing-up and balance a real pole." *IEEE International Conference on Systems, Man and Cybernetics*, Vol. 4, pp.3191 - 3196, 2005

[40] S. Miyata, A. Yanou, H. Nakamura and S. Takehara "Navigation and Path Search for Roving Robot Using Reinforcement Learning." *IEEE International Conference on Networking, Sensing and Control*, pp.480-485, 2009

[41] Pang R. "Multi-UAV formation maneuvering control based on Q-Learning fuzzy controller ." *2nd International Conference on Advanced Computer Control (ICACC)*, Vol. 4, pp.252 - 257, 2010

[42] T.G. Dietterich "The MAXQ Method for Hierarchical Reinforcement Learning." *Proceedings of the 15th International Conference on Machine Learning*, pp. 118-126, 1998

[43] T.G. Dietterich "State Abstraction in MAXQ Hierarchical Reinforcement Learning." *Advances in Neural Information Processing Systems 12 (NIPS-12)*, pp.994-1000, 2000

[44] R. S. Sutton, D. Precup and S. Singh "Intra-Option Learning about Temporary Abstract Actions." *Proceedings of the 15th International Conference on Machine Learning*, pp.556-564, 1998

[45] R. S. Sutton, D. Precup, S. Singh and B.Ravindran "Improved Switching among Temporally Abstract Actions." *Advances in Neural Information Processing Systems 11 (NIPS11)*, pp.1066-1072, 1999

[46] M. Ghavamzadeh and S. Mahadevan "Hierarchically Optimal Average Reward Reinforcement Learning." *Proceedings of the 19th International Conference on Machine Learning*, pp.195-202, 2002

[47] E. Uchibe and K. Doya, "Hierarchical Reinforcement Learning for Multiple Reward Functions." *Journal of the Robotic Society of Japan*, Vol.22, No.1, pp.120-129, 2004

[48] B. Bakker and J. Schmidhuber: "Hierarchical Reinforcement Learning Based on Subgoal Discovery and Subpolicy Specialization." *Proceedings of the Intelligent Autonomous Systems 8*, pp.438-445, 2004

[49] L. Busoniu, R. Babuska and B. D. Schutter "A Comprehensive Survey of Multiagent Reinforcement Learning." *IEEE Transactions on Systemes, Man and Cybernetics*, Vol. 38, No. 2, pp.156-172, 2008

[50] C. F. Touzet "Robot awareness in cooperative mobile robot learning." *Auton. Robots*,vol. 8, no. 1, pp. 8797, 2000

[51] F. Fernandez and L. E. Parker "Learning in large cooperative multi-robot systems." *Int. J. Robot. Autom.*, vol. 16, no. 4, pp. 217226, 2001

Publications

Journal

1). **Syafiq Fauzi Kamarulzaman**, T. Shibuya, S.i Yasunobu, "Substitute Target Learning Based Control System for Control Knowledge Acquisition Within Constrained Environment.", *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol.16 No.3, pp.397-403, 2012.

International conference (reviewed)

1). **Syafiq Fauzi Kamarulzaman**,T. Shibuya, S. Yasunobu, "A Substitute Target Learning-based Inverted Pendulum Swing-Up Control System", *Joint 5th International conference of Soft Computing and Intelligent systems and 11th International Symposium on Advanced Intelligent System*, Japan, SA-D3-3 pp.1-6, 2010.

2). **Syafiq Fauzi Kamarulzaman**, T. Shibuya, S. Yasunobu, "A learning-based control system by knowledge acquisition within constrained environment", *World Congress of International Fuzzy System Association*, Indonesia, FC-104 pp. 1-6, 2011.

3). **Syafiq Fauzi Kamarulzaman**, T. Shibuya, S. Yasunobu, "A Learning-Based Non-Linear Control System with Constraint Consideration", *International Conference of Instrumentation, Control and Information Technology*, Japan, SaA07-02 pp. 1-6, 2011.

4). **Syafiq Fauzi Kamarulzaman**, T.i Shibuya, S. Yasunobu, "A Learning Control System for Rapid Position Control of Aerial Vehicles", *International Conference of Instrumentation, Control and Information Technology*, Japan, pp. 1546-1551, 2012.

5). **Syafiq Fauzi Kamarulzaman**, Seiji Yasunobu, "Cooperative Multi-Knowledge Learning Control System with Obstacle Consideration", *International Conference on Information Processing & Management of Uncertainity in Knowledge-Based System*, France, pp. 505-515, 2014.