Exploring Q-Learning Optimization in Traffic Signal Timing Plan Management

Conference Paper · August 2011

DOI: 10.1109/CICSyN.2011.64 · Source: IEEE Xplore

4 authors, including:

Nurmin Bolong
Universiti Malaysia Sabah (UMS)
141 PUBLICATIONS **1,673** CITATIONS
SEE PROFILE

Soo Siang Yang
Universiti Malaysia Sabah (UMS)
63 PUBLICATIONS **556** CITATIONS
SEE PROFILE

2011 Third International Conference on Computational Intelligence, Communication Systems and Networks

Exploring Q-Learning Optimization in Traffic Signal Timing Plan Management

Yit Kwong Chin[1] Lai Kuan Lee Nurmin Bolong Soo Siang Yang Kenneth Tze Kin Teo[2]
Modeling, Simulation and Computational Algorithm Laboratory
School of Engineering and Information Technology
Universiti Malaysia Sabah
Kota Kinabalu, Malaysia
ykchin@ieee.org[1] ktkteo@ieee.org[2]

***Abstract* — Traffic congestions often occur within the entire traffic network of the urban areas due to the increasing of traffic demands by the outnumbered vehicles on road. The problem may be solved by a good traffic signal timing plan, but unfortunately most of the timing plans available currently are not fully optimized based on the on spot traffic conditions. The incapability of the traffic intersections to learn from their past experiences has cost them the lack of ability to adapt into the dynamic changes of the traffic flow. The proposed Q-learning approach can manage the traffic signal timing plan more effectively via optimization of the traffic flows. Q-learning gains rewards from its past experiences including its future actions to learn from its experience and determine the best possible actions. The proposed learning algorithm shows a good valuable performance that able to improve the traffic signal timing plan for the dynamic traffic flows within a traffic network.**

Keywords - Q-Learning; Traffic Signal Timing Plan; Traffic Flow Control

I. INTRODUCTION

Urban cities usually are covered by a complicated traffic network which is responsible for supporting the demands of the traffic flows in that area. Unfortunately, the traffic demands in the urban area are always high and dynamic. In addition, a fully developed city often limited the availability of the traffic network reconstructions. As a result, traffic congestions occurred around the urban traffic network when the existing traffic network is unable to meet the saturated traffic demands by the on-road vehicles. Traffic lights system is the most common approach used to control the traffic flow within the traffic network to prevent the occurrences of traffic congestions.

As the traffic demands continue to increase, the performance of conventional traffic lights fails to meet the expectation. The traffic lights system needs to evolve more efficiently to learn and adapt towards the dynamic characteristic of the traffic flow. The conventional way of predetermine the traffic lights signals timing plan according to the historical statistics data collected are not good enough to deal with the real traffic flow demands.

The proposed Q-learning control algorithm in this study is focus on the ability of the Q-learning system to learn from its past experiences. The learning ability will assist Q-learning control algorithm to make a better decisions based on its experience, in order to adapt into the dynamic changes of the traffic flow within the urban traffic networks.

II. TRAFFIC SIGNAL TIMING PLAN MANAGEMENT

Traffic lights systems are the most widely used traffic control systems nowadays with different traffic signal planning optimization controls. Traffic lights system works by only allowing a single phase of traffic flow to be passed through the intersections to prevent the intersections from crashed down; as a failure at a single intersection may further lead to the paralyze of the entire traffic network.

The traffic signals timing plans are managed with 3 basic signals which are red signal for stop, green signal for the right to pass the intersection, and amber signal as a gap of time span for the road users to slow down their vehicles at the intersections. A phase is considered completed if the phase has gone through all the 3 signals. After the traffic signal timing plan has circulated all the phases at the intersections, a cycle of traffic signal is completed.

In the management of traffic signal plan, various researches have been carried out throughout these years to improve the performance of the traffic light systems. The earliest traffic signal timing plan is a fixed-time traffic light system where the duration of each traffic signal is set to be fixed. Due to the increasing of vehicles on road, the traffic signal timing system is modified to predetermine the traffic signal timing plan according to the traffic condition collected in the statistic. However, predetermined traffic signal timing plans did not have the ability to react towards the dynamic environment of the traffic networks and failed to adapt towards the dynamic changes of traffic flow. Therefore, artificial intelligence techniques and methods are also introduced into the traffic light systems by different researchers to optimize the traffic flow. Genetic algorithm or evolutionary algorithm is one of the common methods introduced into the traffic lights systems in the other researches [1]. Improvement of the traffic light systems are shown in the research of fuzzy logic control in traffic flow control [2]. Idea such as extending green light period while detecting continuously incoming vehicle flow is implemented in traffic flow control to enhance the performance of traffic light system [3]. Another area of study is using wireless communications between vehicles and traffic control systems to gather information for traffic flow optimization [4]. Reinforcement learning is applied in certain

978-0-7695-4482-3/11 $26.00 © 2011 IEEE
DOI 10.1109/CICSyN.2011.64

studies for the traffic flow control and optimizations in recent years to model and learn the traffic behavior [5, 6, 7]. The proposed Q-learning in this study is one of the reinforcement learning algorithms that are widely used in various fields.

III. Q-LEARNING ALGORITHM

A. Basics of Q-Learning

Q-Learning is one of the common methods available in Reinforcement Learning. Trial-and-error method is used for the exploratory agent to explore in a complex and non-deterministic environment, and then execute (exploitation) the best action based on experience. The experience is based on the reward or penalty that received from the previous trial-and-error action. Reinforcement learning algorithm promised to improve the performance of an agent with the agent's experience.

B. Structure of Q-Learning

Q-Learning is a technique that consists of a few important parts or stages. The main component in Q-learning is the Q-table. Q-table is a matrix table that built up with Q-values, each of the single Q-value represents a value for the states and actions of the algorithm.

The Q-value of the Q-table is evaluated by the following equation,

$$Q(s,a)_i = (1-\alpha)Q(s,a)_{i-1} + \alpha[R(s,a)_i + \gamma_{a'}^{\max} Q(s',a')] \quad (1)$$

where, s = current state
a = action taken in current state
s' = next state
a' = action taken in next state
i = iteration
α = learning rate
γ = discounting factor

Equation (1) shows the agent of Q-Learning will receive a reward or penalty for every action a taken in state, s. At each iteration, the QL agent will select an action with the maximum Q-value at state, s (exploitation), and then evaluate reward function when moving to next state. In order to prevent the algorithm traps inside a certain region only, a greedy probability, ε is introduced. Greedy probability provides a probability to the agent to randomly choose an action from the actions space which does not have the highest Q-value. Therefore, the agent may have a chance to explore to new environment (exploration). A simple flow chart of Q-Learning algorithm is shown in Fig. 1.

The range of the learning rate and discounting factor is in between 0 and 1. If the agent has a higher learning rate, where it is near to 1, then the agent will be more depends on its new knowledge from the learning process than its previous experience. Therefore the agent will have a fast learning speed. The discounting factor actually is to show the importance of the next state. Higher discounting factor, where it is near to 1, meaning the next state is very important and might influence the overall performance of the agent.

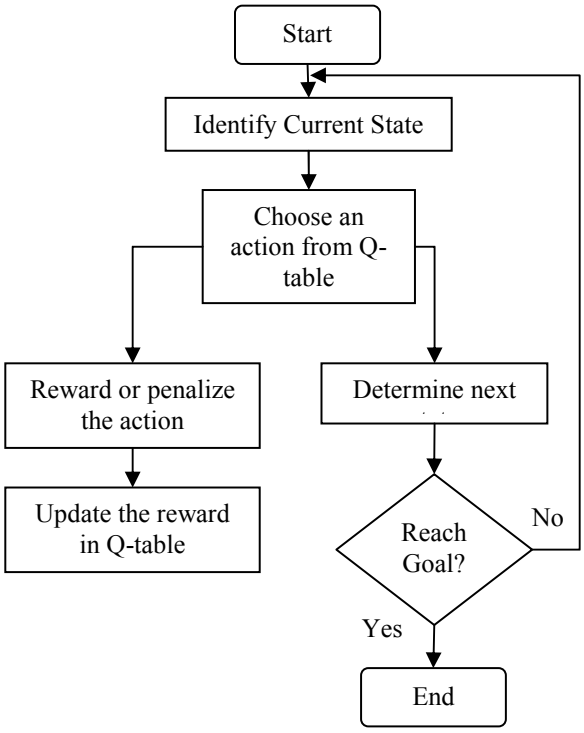

Figure 1. Q-Learning algorithm flow chart.

With all the values obtained in the previous steps, the Q-value for the action a taken in state s can be calculated using (1). Then, this Q-value will be stored in the Q-table. In other words, the Q-table will be defined as the experience of the QL agent. The rewards and penalties of the proposed Q-Learning are set by a set of simple rules.

C. State and Action Definition

Proper defined QL's states and actions are crucial for a QL system to ensure the exploration processes is successfully carried out from a state to another. In this research, the level of the queue length at each phases in the intersection are declared as the states of the QL algorithm. There are 4 levels of queue length defined in this project ranged from no queue length to high queue length. Thus, the combinations of 4 phases at the intersections with 4 levels of queue length result in a total of 256 possible states.

Actions in QL algorithm are the execution acts that enable the algorithm to do exploration within the states of QL. Each action will lead the algorithm from the current state to the other states. The actions available for the developed QL algorithm in this research are the green signal

distributions, which are 1 seconds and 5 seconds of green signals. The green signal actions will be stored and distributed among the 4 traffic phases once they are chosen throughout the exploration of the QL algorithm for the exploitation purpose. In this study, no 0 green signals are defined as penalties will be given when both the actions appear to be the wrong decisions. As a results, the algorithm will proceed to explore the suitable green signals distribution of others traffic phases. The penalties and rewards of each available action are evaluated in the reward and penalty functions.

D. Reward and Penalty Functions.

Q-Learning determines its best action based on the reward returned by that particular action. In shorts, the highest reward returning action is the best results. The proper rules or policies to reward and penalize the actions are crucial to help the Q-Learning to get the most optimum action. The proposed Q-Learning based traffic flow control and optimization system are responsible to archive the lowest vehicles in queue at the intersection at any time. Thus, the rewarding and penalties of the actions must help to decide which action is the best to produce the least number of vehicles in queue at the intersection.

Rewards are given to the actions when there are vehicles in the queue at the intersection. This rule causes problem when the selected action caused too much green signal duration for the traffic condition, as green signal is still on where there is no more vehicle waiting at the intersection. The solution for the problem is to set a penalty for the actions. The penalties of the actions are given when there are occurrences of green signal duration wasted because of the action.

During the oversaturated traffic condition, traffic lights system's efficiency is significantly drop as the traffic lights system cannot support too many traffic demands. Due to the continuously incoming vehicles at a heavy traffic intersection, the green signal duration will be larger. Therefore, if the QL algorithm tends to clear a particular traffic phases, it will cause more vehicles to accumulate at the other traffic phases and lengthen the vehicle queue. As a result, penalty will be introduced on the action when too much green signal are allocated upon a single traffic phase. This step is to compensate and optimize the average waiting time of the traffic users at the intersection during the saturated traffic condition. A penalty factor will be introduced into the algorithm, where the penalty factor of each traffic phase will increase for every distributed green signal. As time goes on, the penalty factor will become significantly large to warn the algorithm about the traffic phase getting too much green signals.

However, a stopping criterion is defined in the algorithm to indicate the objective accomplished in the QL algorithm. The QL will stop when all the traffic phases are distributed with optimum traffic signal timing plan as well as no more queue length at the intersection.

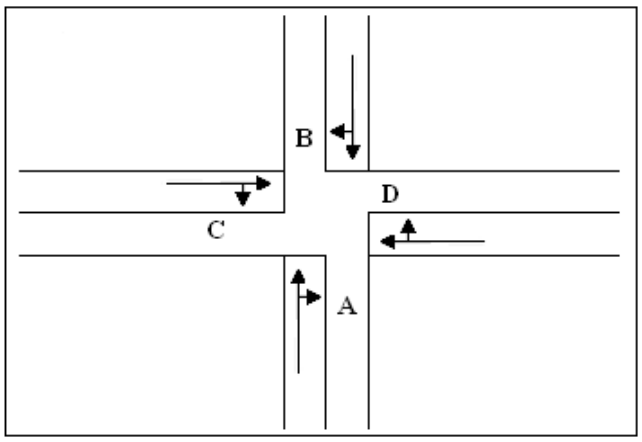

Figure 2. 4-way traffic intersection with 4 phases.

The reward and penalty of the actions are continuously updated into the Q-table for the purpose of exploration and exploitation in the future.

IV. SIMULATIONS

A. Traffic Intersection

In this study, a 4-way intersection in front of University Malaysia Sabah (UMS) is used as the model of the study which consists of 4 phases.

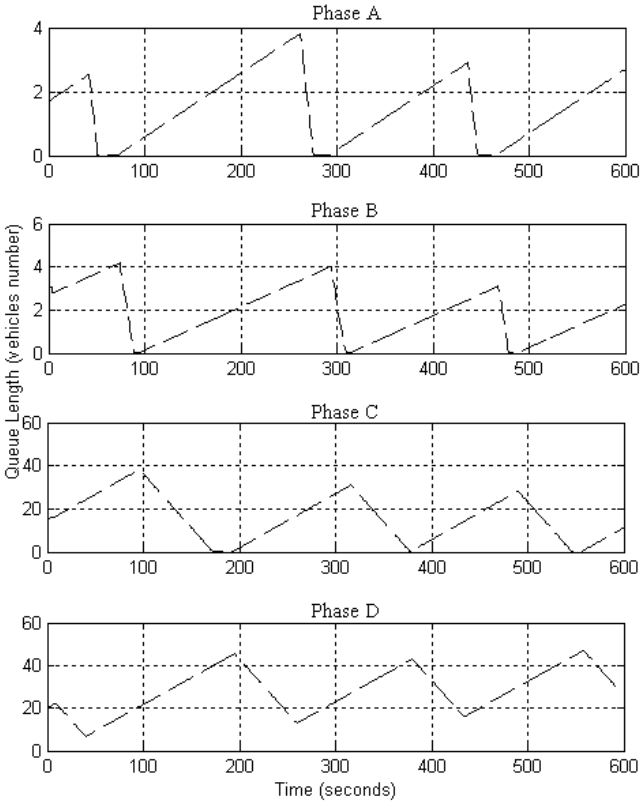

Figure 3. Simulation results of QLTSTM at UMS intersection.

Phases are the sequence of turns for the traffic signals timing plans to allow only certain phase of traffic flows to pass through the intersection at a particular time [8]. The 4-way intersection shown in Fig. 2 is labeled with 4 phases, which are phase A, B, C and D. The setting of phases is important to prevent crashes of vehicles at the intersections, and ensure the traffic scheduling or traffic signals timing plan of the system is efficient.

B. Description of Traffic Intersection

4-way intersection is chosen as the simulation platform. The data collected at the intersection is used to test the performance of the developed QL traffic signal timing plan management system (QLTSTM). The results for the QLTSTM system is shown in Fig. 3.

Fig. 3 shows the simulation results of QLTSTM at UMS intersection. Traffic phase C and D are experiecing more traffic flow than traffic phase A and B since more incoming vehicles at the main raod phase C and D. The simulation has been run for 600 sceonds to carry out the analysis on the developed QLTSTM system. Simulation results on different phases show that QLTSTM system is able to determine a suitable timing plan for the intersection as all the traffic phases manage to maintain a low level of vehicle queue length.

The increasing slope of the graphs indicates the traffic phase is undergoing a red signal waiting time as the vehicles started to accumulate while the decreasing part of the graphs show the green signal activated for releasing the vehicles in queue to pass the intersection. Besides, observation on the green signal of different graphs indicate that only one traffic phase is given the green signal at a particular moment.

In order to test the performance of the developed QLTSTM system, three different situations should be tested with various traffic conditions. First, the simulation of an increasing traffic demand towards the QLTSTM system has been carried out. Then, the response of the QLTSTM on the decreasing traffic demand could be analyzed through the simulation. Lastly, the simulation to test the adaptivity of the QLTSTM system in the dynamic changes due to the traffic environment has been carried out.

V. RESULTS AND DISCUSSIONS

A. Results of Simulations

The simulations have been carried out for 3600 seconds in various cases to evaluate the long term performance of the QLTSTM system. In the first case, the traffic flow input have been set to an average low traffic flow input for 15 minutes, then the traffic flow input started to increase to a saturated traffic flow input after. The purpose of running average low input for 15 minutes of simulation time is to let the QLTSTM system reach the steady state response for the average low traffic flow input before the high saturated traffic flow disturbance has been introduced into the simulation. This simulation represented the situation where

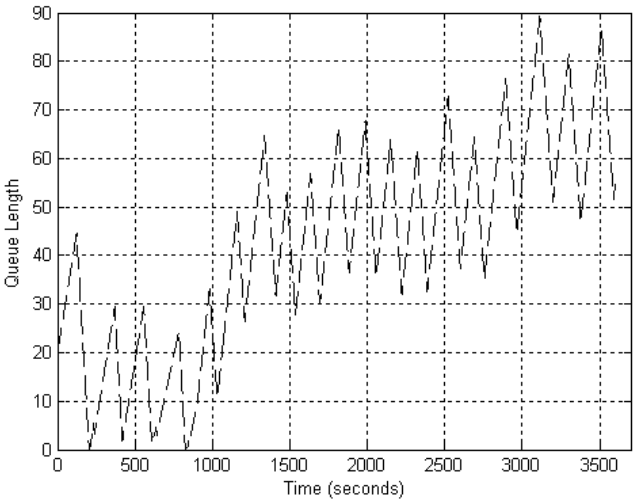

Figure 4. Phase D simulation results for situation 1.

the traffic condition started to shift into the peak hour of the day from an average traffic situation.

Refer to Fig. 4, the results of the simulation for traffic phase D is shown. Result on traffic phase D is analyzed and observed because traffic phase D is experiencing the most significant changes in traffic conditon at the UMS intersection. The results show that QLTSTM system manage to handle the sudden increase of the traffic flow in.

At the beginning of the simulation, QLTSTM able to reduce the vehicles in queue effectively, but at the later part of the simulation, the queue length at the traffic phase increase rapidly and start to reach a steady state and leave the traffic phase with average more than 35 vehicles at the intersection after each traffic cycle.

Another traffic situation has been tested after the peak hour. The simulation started with the oversaturated traffic flow for 15 minutes and then continued with the disturbance from the average low level of trafiic flow in. Fig. 5 shows

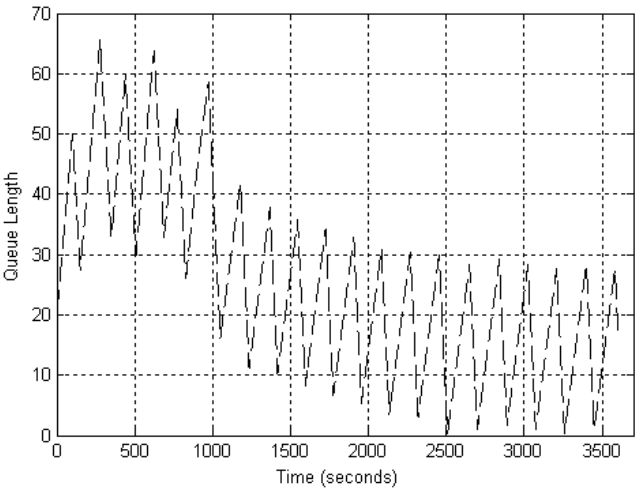

Figure 5. Phase D simulation results for situation 2.

the results from the traffic phase D as it is the main interest in this observation. The performance of the QLTSTM system has been studied based on the observation of Fig. 5.

For the first 15 minutes, the traffic situation is oversaturated as the maximum vehicles left at the intersection reached an average value of 65 vehicles in queue. After 15 minutes, when the traffic flow in has been backed into an average low level, QLTSTM system managed to start reducing the vehicles in queue at the intersection. At the later stage of the simulation, QLTSTM system able to reduce the queue length at the intersection to a minimum level.

The third or final case of this study is to imitate the random dynamic changes of the traffic environment, and the QLTSTM is tested with the model. The traffic condition has benn changing throughout the simulation for different traffic phases. Two traffic phases have been shown in this simulation for a better observation on the performance of the QLTSTM, which is the two main heavy traffic phases. Fig. 6 consists of two graphs, first graph is the results for traffic phase C and the second is the traffic phase D. In the first 500 seconds of the simulation, both traffic phases have been inserted with average traffic flow, and the QLTSTM performed expectedly well for both traffic phases.

Then from 500 seconds to 1000 seconds, traffic phase C is experiencing with the saturated traffic flow while traffic phase D still having the same traffic condition. The graph in Fig. 6 indicated that QLTSTM system, increased the green signal duration for traffic phase C to release more vehicles in queue. From time 1000 seconds to 1500 seconds, the traffic input changed back to average low for both traffic phases. The observation on the results started to get interested during the simulation time 1500 seconds to 2500 seconds, as traffic phase D has been fed with a heavy traffic flow. This heavy trafiic flow caused a massive amount of vehicles into the intersections during that period and the queue length of traffic phase D to reach a maximum point of 58 vehicles. However, QLTSTM system still able to release most of the vehicles at that period in 3 traffic cycles. Last part of the simulation has been carried out with the traffic phases via average low traffic flow. The simulation performed accordingly to the expectation of the developed QLTSTM system where it was able to maintain a minimum level of vehicles in queue at both traffic phases.

B. Discussions

During simulation of situation 1 where the traffic condition became the peak hour of the day, QLTSTM system managed to adapt itself into the situation providing optimal green signal duration. Although the queue length of the traffic phase D did not maintain at the minimal level as before the oversaturated traffic condition, QLTSTM still able to release most of the vehicles waiting at the intersection.

The optimization of QLTSTM system was clearly seen in this case where it compromised the green signal duration

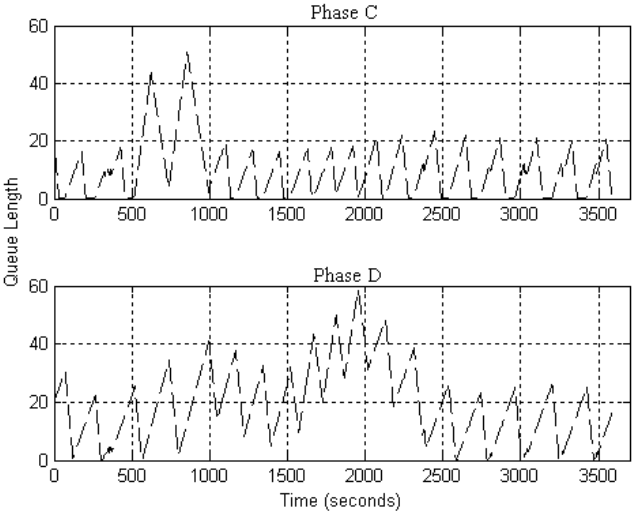

Figure 6. Phase C and D results for situation 3.

for traffic phase D to reduce the average waiting time of other traffic phases. The compromises of QLTSTM were observed through Fig. 4, QLTSTM did not continue to distribute green signal for traffic phase D. It happened since the penalties in the reward function of the QL algorithm which restricts itself to distribute too much green signal duration to the certain traffic phase. However, QLTSTM had successfully distributed the maximum green signal duration available for the traffic phases, which is about 50 seconds for the optimization purpose of the traffic flow control.

The results in the simulation of situation 2 have shown QLTSTM system capability to react fast towards the traffic condition. In the first 15 minutes of the simulation, heavy traffic flow had burden the traffic phase where it accumulated a massive amount of vehicles, but QLTSTM system successfully maintained its level of vehicles in queue and reacted fast once the traffic flow changes. Based on Fig. 5, QLTSTM system used 6 traffic cycles within 1000 seconds and 2000 seconds to release most of the vehicles in queue at the intersection right after it detected the changes of the traffic condition.

The last study of the developed QLTSTM system focused on the adaptability of the system to the dynamic environment of the traffic networks. Throughout the simulations, various combinations of traffic flow input have been implemented to evaluate the QLTSTM system. The effect of heavy traffic flow towards other traffic phases was investigated based on the observation of Fig. 6. During 500 seconds to 1000 seconds, traffic phase C experienced a sudden heavy traffic flow changes, the vehicles in queue had a sudden rise to 40 vehicles above and more vehicles had caused the QLTSTM system to allocate more green signals duration for traffic phase C. The effect of the action in QLTSTM system had lead to a slight increase in the level of vehicles queue length at traffic phase D. It was

understandable since more green signals allocated to a particular traffic phase will increase the red signals duration of other traffic phase. Thus, the ability of the QLTSTM system can evaluate their actions towards their rewards and penalties through the reward functions are evaluated and assessed.

VI. CONCLUSIONS

In this paper, the studies on the traffic flow control systems have been carried out. The developed Q-learning based traffic signal timing plan management systems had the ability to perform well in various traffic environments.

The simulation results verified the QLTSTM system had the ability for compromising to have less green signal durations for the purpose of reducing the average waiting time and queue length of the other traffic phases. QLTSTM also able to react fast towards the changes of traffic flow input. In addition, the QLTSTM system can adapt itself in the dynamic changes of the incoming traffic flow and also the vehicles in queue.

Q-learning algorithms' ability to explore itself in the dynamic traffic flows and decide its best actions based on its experience has shown a good performance in the traffic signal timing plan management system. The ability of Q-learning encouraged the traffic signal timing plan system to be able to adapt into the dynamic changes of traffic flow system. The simulations of this study had shown that Q-learning is a suitable method or technique to be implemented into the traffic flow control and optimization of urban traffic network system.

ACKNOWLEGEMENT

The authors would like to acknowledge the funding assistance of the Ministry of Higher Education of Malaysia (MoHE) under Fundamental Research Grant Schemes (FRGS), grant No. FRG0105-TK-1/2007 and FRG0220-TK-1/2010, University Postgraduate Research Scholarship Scheme (PGD) by Ministry of Science, Technology and Innovation of Malaysia (MOSTI).

REFERENCES

[1] Fitsum Teklu, Agachai Sumalee, David Watling. "A Genetic Algorithm Approach for Optimizing Traffic Control Signals Considering Routing.", Institute for Transport Studies, University of Leeds. 2006.

[2] Ehsan Azimirad, Naser Pariz, and M.Bagher Naghibi Sistani, "A Novel Fuzzy Model and Control of Single Intersection at Urban Traffic Network",IEEE Systems Journal,Vol. 4, No. 1, March 2010, pp107-111.

[3] Kok Khiang Tan, Marzuki Khalid, Rubiyah Yusof. "Intelligent Traffic Lights Control By Fuzzy Logic." Malaysian Journal of Computer Science. 1996.

[4] Victor Gradinescu, Cristian Gorgorin, Raluca Diaconescu, Velentin Cristea. "Adaptive Traffic Light Using Car-to-Car Communication." University Bucharest Computer Science Department. 2007.

[5] I. Arel, C.Liu, T.Urbanik, and A.G. Kohls, "Reinforcement learning-based Multi-agent system for network traffic signal control.", IET Intelligent Transport Systems, Vol. 4, Iss. 2, 2010, pp. 128-135.

[6] Liu Zhi-Yong, and Ma Feng-wei, "On-line Reinforcement Learning Control for Urban Traffic Signals.",Proceedings of the 26th Chinese Control Conference, 2007, pp. 34 – 37.

[7] P. G. Balaji, X. German, and D. Srinivasan, "Urban Traffic Signal Control using Reinforcement Learning Agents", IET Intelligent Transport Systems, Vol. 4, Iss. 3, 2010, pp. 177 – 188.

[8] Nicholas j. Garber, Lester A.Hoel. "Traffic and Highway Engineering 3rd Ed. Thomson.", 2002.