Q-Learning Based Traffic Optimization in Management of Signal Timing Plan

Yit Kwong Chin, Nurmin Bolong, Aroland Kiring, Soo Siang Yang, Kenneth Tze Kin Teo
Modelling, Simulation and Computing Laboratory,
School of Engineering and Information Technology,
Universiti Malaysia Sabah, Kota Kinabalu, Malaysia.
msclab@ums.edu.my, ktkteo@ieee.org

***Abstract* - Occurrences of traffic congestions within the urban traffic network are increasing in a rapid rate due to the rising traffic demands of the outnumbered vehicles on road. The effectiveness of management from traffic signal timing planner is the key solution to solve the traffic congestions, but unfortunately the current traffic light signal system is not fully optimized based on the dynamic traffic conditions on the road. Adaptable traffic signal timing plan system with ability to learn from their past experiences is needed to overcome the dynamic changes of the urban traffic network. The ability of Q-learning to prospect gains from future actions and obtain rewards from its past experiences allows Q-learning to improve its decisions for the best possible actions. A good valuable performance has been shown by the proposed learning algorithm that able to improve the traffic signal timing plan for the dynamic traffic flows within a traffic network.**

Keywords - Q-learning, signal timing plan, traffic control, learning algorithm

I. INTRODUCTION

The demands of the traffic flows in the urban area are usually supported by a complicated traffic network which covers the whole urban cities. Unfortunately, in the urban area, there are always high traffic demands with dynamic traffic conditions. In addition, reconstructions of the traffic road to cope with the high traffic demands are not an option for a fully developed city with limited availability of landscapes. When the existing traffic network is unable to meet the saturated traffic demands by the on-road vehicles, traffic congestions occurred around the urban traffic network. The most common solution for the traffic congestions problem is the implementation of traffic lights system to control the traffic flow within the traffic network.

The breakdown of the conventional traffic light~~s~~ system begins when it fails to fulfill the rising traffic demands of the traffic network. Therefore, a breakthrough evolution of the traffic lights system is needed to learn and adapt towards the dynamic characteristic of the traffic flow. The conventional method of predetermine the traffic lights signals timing plan based on the historical traffic statistics data is insufficient to handle the actual traffic flow demands.

The ability of the Q-learning system to learn from its past experiences is focused in the study of traffic signal timing plan management system. The experiences learnt by Q-learning control algorithm from its past actions will assist the algorithm to make better decisions in future for its adaption into the dynamic changes of the traffic flow within the urban traffic networks.

II. REVIEW OF TRAFFIC SIGNAL TIMING PLAN

Traffic lights systems with different traffic signal planning optimization approaches are being widely used to control the traffic flow nowadays. The system functions by coordinating the signal timing plan to ensure every phases of the traffic flow has the permission to pass through the intersection and preventing the intersection from crashed down. The paralysis of the entire traffic network can be caused by a failure at a single intersection.

Red, green and amber signals are the 3 basic signals in the traffic signal timing plans which signalize the stoppage of vehicles at intersection, permission to pass through intersection, and warning for slow down before the intersection respectively. After all the 3 signals have been given to a link in the intersection, a phase is considered completed. A cycle of traffic signal is completed~~,~~ after the traffic signal timing plan has circulated all the phases at the intersections.

Various studies have been carried out throughout these years for the enhancement of the traffic light systems through the management of the traffic signal plan. Fixed-time traffic light system is one of the primitive approaches in traffic signal timing plan, where the duration of each traffic signal is determined beforehand. The setup of the traffic signal timing system is based on the historical statistic of the traffic condition. However, this method will not allow the traffic signal timing plan to react towards the dynamic changes of the traffic condition in the variable environment of traffic networks. Therefore, artificial intelligence techniques consists of learning ability has been proposed in the evolution of the traffic light systems. Different researchers have chosen variant types of artificial

intelligence methods for the optimization of the traffic flow. Genetic algorithm or evolutionary algorithm is one of the common methods introduced into the traffic control systems. Consideration of routing of traffic flow using genetic algorithm has shown some improvement in the traffic control [1]. Fuzzy logic control is also being implemented into the traffic light systems for better control of traffic flow [2]. Enhancement of the performance of traffic light system is done with idea such as extending green light period while detecting continuously incoming vehicle flow [3]. Another approach to improve the traffic control is using wireless communications between vehicles and traffic control systems to gather information for traffic flow optimization [4]. Reinforcement learning is applied in certain 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

Reinforcement Learning is an algorithm that is able to improve itself through the learning process of the past. Q-Learning is one of the common methods available in reinforcement learning. In Q-learning, the exploratory agent explores in a complex and non-deterministic environment via trial-and-error approach, and then executes (exploitation) the best action based on the experience [8, 9]. Rewards or penalties will be gained by the previous trial-and error actions and stored in the memories as experiences for future references. Q-learning algorithm promised to improve the performance of an agent with the experience gained in the past.

B. Structure of Q-learning

Q-table is the main component of the Q-learning algorithm. The Q-table is a matrix table where various information extracted from the traffic plan is being stored. Each element in the Q-table is identified as Q-values, which represents a value for every single states and actions pair. The algorithm evaluates Q-value of the Q-table by (1).

The agents of Q-learning will receive a reward or penalty for every action a taken in state s with the evaluation from (1). In every iteration steps, the maximum Q-value at state s will be selected by the Q-learning agent (exploitation). Then, evaluation from the reward function for that action will be stored in the Q-table when the algorithm moves to the next state [10].

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

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

C. ε -Greedy Selection

Actions of the Q-learning algorithm are determined by their value of rewards or the nature of randomness. Selection by the value of rewards is a simple task of locating the maximum rewards available from the action lists. However, the selection process should be conducted in random. The random selection can be used to prevent the selection process being trapped at local maximum or minimum point.

This is where the greedy probability, ε is introduced into the algorithm. Greedy probability, ε allows the Q-learning to have a chance of randomly choose an action from the actions space which does not consist the highest Q-value [11]. The introduction of the greedy probability ensures the agent having a chance to explore into the new environment (exploration). Without the greedy probability, ε, Q-learning process will lack the tendency of exploration in the new environment which will end up with missing the chance to get in touch with a new experience with larger rewards. However, if the greedy probability, ε is too large, the stability of Q-learning will decrease as the exploration of the algorithm will hinder itself from settling down in the known environment with experience.

D. Learning Rate and Discount Rate

Learning rate and discount factor are the two elements that influence the convergence speed of the algorithm. The learning rate and discounting factor is ranged in between 0 and 1. Learning rate of the Q-learning algorithm determines the importance of the newly acquired experience. When the agent acquires a higher learning rate, where it is near to 1, then new knowledge from the learning process will have more influence towards the agent than its previous experience. If the learning rate is set to lower value or 0, then the Q-learning will not acknowledge the newly gained experience, and act upon solely on the past gained experience. This will inhibit the Q-learning algorithm from learning the environment. Thus, a suitable learning rate shall be determined for the agent to be able to learn faster in the new environment.

The discounting factor is capable of showing the degree of importance of the next state. Higher discounting factor when the value is near to 1 means the future gain prospected by the agent is more important. It might lead to faster convergence of Q-table and thus influence the overall performance of the agent. However, a high discounting factor will influence the Q-learning from

looking at the importance of the current experience. Therefore, discounting factor has to be optimum to let Q-learning algorithm has a balance focus on both the short-term rewards from the current experience and also the long-term benefit from the future prospect gain.

E. Flow Chart of Q-learning

The flow chat of the Q-learning algorithm is shown in Fig. 1. The Q-learning algorithm will start with identifying the current state s from the input. After the identification of the state s , an action will be chosen from the action list, either by searching for the maximum rewards or by random if the greedy probability ε is triggered. With all the values initialized in the previous steps, the Q-value for the action a taken in state s is calculated using (1). Q-table will then store the Q-value. In other words, the experience of the QL agent is defined inside the Q-table. The rewards and penalties of the proposed Q-Learning are evaluated by a set of simple rules of reward functions. Following the evaluating reward and updating the Q-table, the next state s' for the Q-learning algorithm will be determined after the selected of the action a is executed. Then the stopping criteria for the Q-learning algorithm will be checked with the determined next state, s'. If the next state s' is the final goal of the Q-learning, then the process will be ended, else the next state s' will become the current state s for another new iteration. The process will be continued until the goal or stopping criteria are fulfilled.

F. States and Actions

The environment model of the Q-learning algorithm is defined from the definition of the states and actions. A proper defined states and actions are crucial for a Q-learning system to ensure the exploration process can be successfully implemented throughout all the possible states. The level of queue length at each traffic phases in the intersection is chosen as the Q-learning algorithm states. The states of the Q-learning are categorized into 4 levels of queue length in this study from no queue length to high queue length. Thus, there are total of 256 possible states combination from the permutation combination of 4 phases at the intersections with 4 levels of queue length each.

The environment of Q-learning is modelled by states. In order to accomplish the exploration within the environment, execution acts that in charge of the exploration are needed. The acts are defined as actions and responsible for leading the algorithm from the current state to the other states. In this study, the green signal distributions of 1 second and 5 seconds are the action of the proposed Q-learning algorithm. The green signals are chosen by the Q-learning algorithm, and then being stored in the memory and distributed to the traffic phases after the Q-learning algorithm reaches its goal. Penalties will be given when both the actions appear to be the wrong decisions and hence no zero value of green signals is defined. The algorithm will proceed to explore the suitable green signals distribution of others traffic phases when there are certain penalties received. The penalties and rewards of each available action are evaluated in the reward and penalty functions.

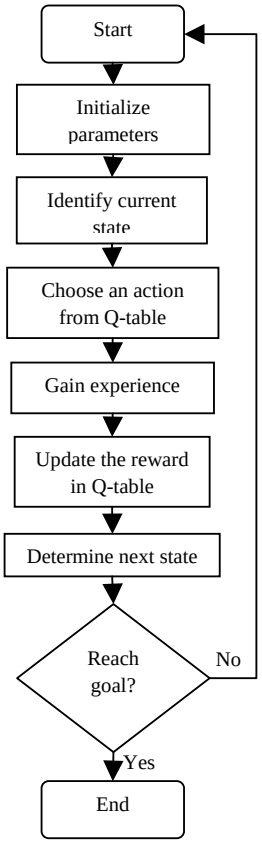

Figure 1. Q-Learning algorithm flow chart.

G. Rewards and Penalties Function

In Q-learning algorithm, each of the actions will return a certain rewards, meaning that the best action will acquire the highest rewards in the process. In order to let Q-learning decides the most optimum action without errors, proper rules or policies to reward and penalize the actions have to be carefully carried out. In traffic flow control and optimization system, the best outcome is when the system is able to archive the lowest vehicles in queue at the intersection at anytime. Thus, Q-learning algorithm must have the ability to decide the best traffic signal timing plan to produce the least number of vehicles in queue at the intersection. Appropriate rewards and penalties functions will be formed to ensure the best optimum timing plan can be built. In the proposed Q-learning algorithm, the actions are rewarded when green signal is distributed for the vehicles in the queue at the intersection. If the action is allocating unused green signal to the phase without any

vehicles waiting, a penalty will be given to the action. Theoretically, it is a waste for a longer green signal to be distributed to the phase with lesser vehicles.

During the oversaturated traffic condition, continuously incoming vehicles at a heavy traffic intersection will lengthen the green signal duration. Therefore, the Q-learning algorithm will tend to distribute longer green signal to the traffic phase, as there will always be vehicles at the intersection due to the heavy traffic flow to clear the particular traffic phase. But that decision will cause more vehicles to accumulate at the other traffic phases because of the long waiting time, and this will lengthen the vehicle queue. As a result, the action will be fined or penalized when too much green signal is allocated upon a single traffic phase. The purpose of this second penalty is to compensate and optimize the average waiting time of all the traffic users at the intersection during the saturated traffic condition. The penalty works by introducing a penalty factor into the algorithm, where the penalty factor of each traffic phase will increase for every distributed green signal. The penalty factor will become significantly large to act as a warning about the traffic phase getting too much green signals as time goes on.

However, a stopping criterion has to be set in the algorithm to indicate the accomplishment of objectives in the Q-learning algorithm. Q-learning algorithm 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.

The reward results from the evaluation of the actions are continuously updated into the Q-table as Q-value for the purpose of exploration and exploitation in the future.

IV. SIMULATONS

A. Traffic Intersection

A 4-way intersection in front of University Malaysia Sabah (UMS) is used as the model in this study, which consists of 4 phases. Phases are the sequence of the traffic signals to allow only certain traffic flows to pass through the intersection at a particular time in the traffic signal timing plan management [12]. Fig. 2 shows the 4-way intersection which is labeled with 4 phases, namely phase A, B, C and D respectively. The efficiency of the traffic scheduling or traffic signals timing plan is ensured by the traffic lights setting of all the phases. Besides that, prevention of vehicle crashes at the intersections is carried out with the proper setting of the phase sequence.

B. Description of Traffic Intersection

The 4-way intersection is chosen as the simulation platform with the collected traffic data at the study site. Performance of the developed QL traffic signal timing plan management system (QLTSTM) is tested using the data collected. The results of the simulated QLTSTM system are shown in Fig. 3.

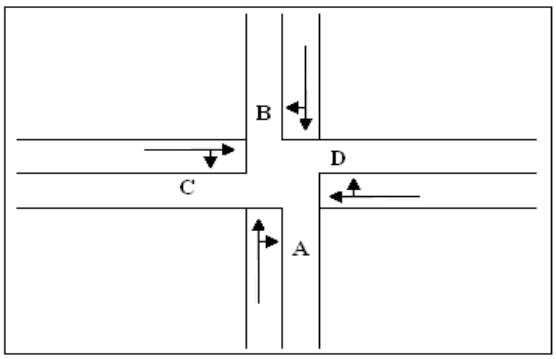

Figure 2. 4-way traffic intersection with 4 phases

Fig. 3 shows the simulation results of QLTSTM at UMS intersection. The developed QLTSTM system selected simulation time of 600 seconds to carry out the analysis for its performance. It can be observed that due to the heavy incoming vehicles at the main road, phase C and D experiencing more traffic flow than traffic phase A and B. Simulation results show that all the traffic phases manage to maintain a low level of vehicle queue length. It is proven that QLTSTM system is able to determine a suitable timing plan for the intersection.

During the traffic phase undergoing a red signal, the vehicles in queue will experience waiting time and start to accumulate which cause the slope of the graph to increase. Meanwhile, the decreasing part of the graphs show the green signal is activated for releasing the vehicles in queue to pass the intersection. From the graphs, observations show that each phases is undergoing their green signal at different timing, this indicates that only one traffic phase is given the green signal at a particular moment.

Three different situations should be tested with various traffic conditions in order to test the performance of the developed QLTSTM system. First, QLTSTM is tested and simulated with an increasing traffic demand. Then, the response of the QLTSTM on the decreasing traffic demand is also examined and analyzed through the simulation. Finally, a simulation of the QLTSTM system in the dynamic changes due to the traffic environment has been carried out to test the adaptability and robustness of the system. The results of the simulation are shown in the next section.

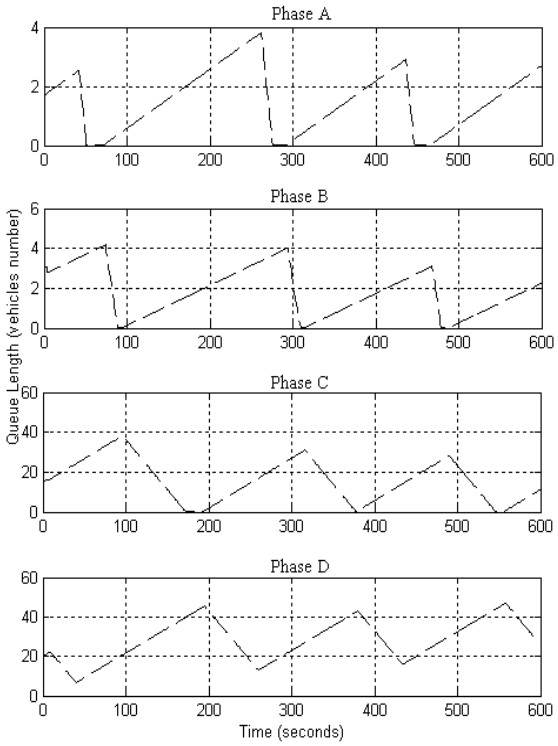

Figure 3. Simulation results of QLTSTM at UMS intersection.

V. RESULTS AND DISCUSSIONS

A. Results of Case I Simulations

Simulations of the QLTSTM have been carried out for 3600 seconds in various cases to evaluate the long term performance of the QLTSTM system. In Case I, approximately 900 seconds of an average low traffic flow is fed into the QLTSTM simulator. Then, the traffic incoming flow is started to increase until the end of the simulation. The collected practical data of the incoming traffic flow indicates the traffic condition before and during the peak hour of the day. During the first 900 seconds, the traffic condition is classified as non peak hour. The incoming flow is considered not heavy before reaching the steady state condition. After that, the incoming flow started to increase and the traffic becomes congested during the peak hour. Fig. 4 shows the simulation result of traffic phase D in Case I. Traffic phase D is chosen for analysis and discussion since it experiences the most significant changes in traffic condition at the UMS intersection. The result shows that the sudden increase of the incoming traffic flow is able to be managed by the proposed QLTSTM system. At the beginning of the simulation, QLTSTM releases most of the vehicles in queue effectively. However, it can be observed that the queue length at the traffic phase increases due to the sudden rising of the incoming flow Nevertheless, QLTSTM start to adapt in the situation by releasing more vehicles and the system manages to maintain approximately 30-60 vehicles in queue at the traffic intersection after each traffic cycle.

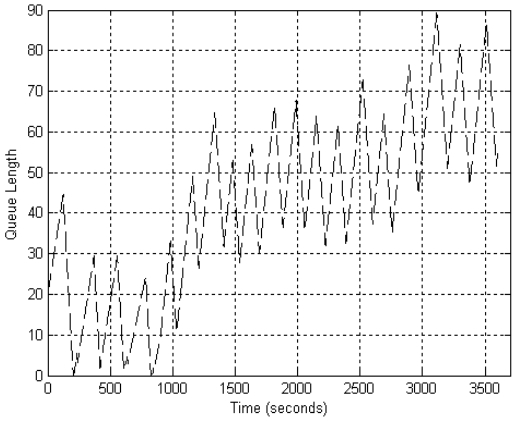

Figure 4. Simulation result of Phase D in Case I.

B. Results of Case II Simulations

Case II is referring to the traffic condition after the peak hour. The simulation begins with 900 seconds of oversaturated traffic flow and then the incoming traffic flow is gradually reduced. Fig. 5 shows the results of traffic phase D in Case II. It can be noticed that the traffic situation is oversaturated for the first 900 seconds where the maximum queue length has reached a maximum value of 65 vehicles. After the oversaturated situation, the incoming traffic flow has been reduced. The results show that the traffic condition at the intersection is eased by the QLTSTM at the later stage of the simulation, as the proposed algorithm is able to reduce the queue length at the intersection to a minimum level. At the end of every traffic phase after 1500 seconds, the vehicles in queue are reduced under 10 vehicles.

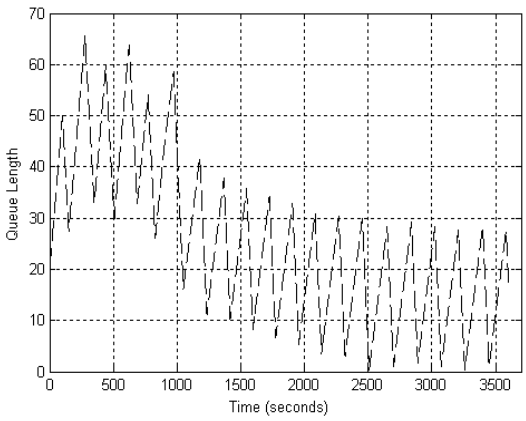

Figure 5. Simulation result of Phase D in Case II.

C. Results of Case III Simulations

Case III involves the random dynamic changes of the traffic environment and the QLTSTM is tested under this situation. Throughout the simulation, the traffic condition is varies at different traffic phases. Two heavy traffic phases are compared for a better observation on the performance of the QLTSTM system. Fig. 6 shows the results of traffic phase C and traffic phase D respectively. Both traffic phases are introduced with average traffic flow during the first 500 seconds of the simulation, and the QLTSTM performed expectedly well.

During 500 seconds to 1000 seconds, traffic phase D still maintaining the same traffic condition whereas traffic phase C is experiencing the saturated traffic flow. The QLTSTM system increases the green signal duration of traffic phase C to release more vehicles in queue as shown in Fig. 6. From time 1000 seconds to 1500 seconds, the incoming traffic reduces to average low level in both traffic phases. During the simulation time of 1500 seconds to 2500 seconds, traffic phase D has been fed with a heavy traffic flow, indicating the possibility of the begin of traffic congestion. Due to rapid increase in the incoming flow, the queue length of traffic phase D reaches 58 vehicles. However, most of the vehicles are able to be released after 3 traffic cycles via QLTSTM system's fast response towards the situation. Last part of the simulation has been carried out with the traffic phases under average low traffic flow, and the results show that QLTSTM manage to maintain the level of vehicles in queue after various traffic condition. The developed QLTSTM system's performance meets the expectation since it is able to maintain a minimum level of vehicles in queue at both traffic phases in the simulation.

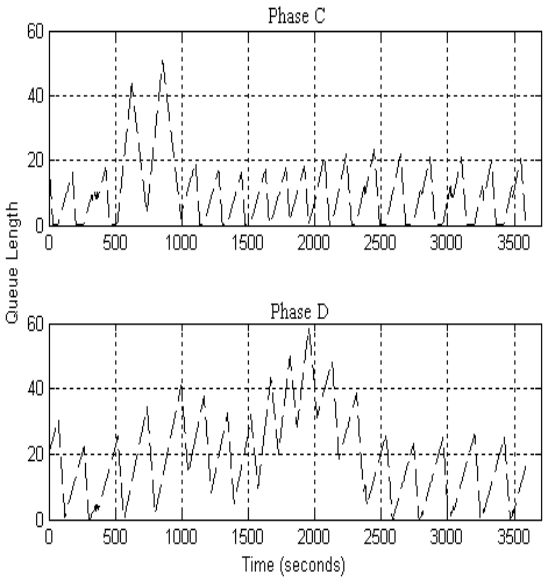

Figure 6. Simulation results of Phase C and D in Case III.

D. Discussions

During simulation of Case I where the traffic condition refer to the peak hour of the day, optimal green signal duration gained from the QLTSTM system is able to optimize and control the traffic situation. QLTSTM still release most of the vehicles waiting at the intersection, even though the queue length of the traffic phase D cannot maintain at the minimal level as before the oversaturated traffic condition.

QLTSTM system determines the green signal duration for traffic phase D in this situation to reduce the average waiting time of other traffic phases for the aim of traffic flow optimization. Based on the observation in Fig. 4, QLTSTM does not continuously to distribute green signal for traffic phase D as the result of its ability to compromise own interest. The system only allocates 50 seconds of green time for traffic phase D in each traffic cycle, even there are still vehicle in queue. The penalties in the reward function of the Q-learning algorithm restrict itself to distribute too much green signal duration to certain traffic phase. The ability of QLTSTM to defend the global benefits instead of local interest is verified throughout the simulation result.

QLTSTM system's capability to react fast towards the traffic condition is shown in the results of Case II. Although during the first 900 seconds, heavy traffic flow has burdened the traffic phases where a massive amount of vehicles is accumulated, QLTSTM system still successfully maintains its level of vehicles in queue and reacts fast to the traffic flow changes. Based on Fig. 5, 6 traffic cycles are used by QLTSTM system within 1000 seconds and 2000 seconds to release most of the vehicles in queue at the intersection right after it detects the changes of the traffic condition.

The adaptability of the system to the dynamic environment of the traffic networks is focused in the final study of the simulation. Various combinations of incoming traffic flow have been implemented to evaluate the QLTSTM system. Based on the observation of Fig. 6, investigation on the effect of heavy traffic flow towards other traffic phases has been carried out. A sudden heavy traffic flow changes applied to traffic phase C during 500 seconds to 1000 seconds where the vehicle in queue has a sudden rise over 40 vehicles. QLTSTM system has to allocate more green signals duration for traffic phase C in order to cope with the traffic demands. Vehicle queue length at traffic phase D has experienced a slight increase because of the action of QLTSTM in phase C. It is reasonable since the red signals duration of other traffic phases will increase as longer green signal is allocated to a particular traffic phase. Thus, the ability of the QLTSTM system to reward and penalize towards the action via the reward function has been evaluated and assessed.

VI. CONCLUSIONS

In this paper, studies have been carried out on the traffic flow control systems. The developed Q-learning based traffic signal timing plan management system has shown its performance through the simulations and the ability to perform well in various traffic environments are proven.

The simulation results verify the QLTSTM system has the ability for optimizing the traffic phase to utilize less green signal durations. With the purpose of reducing the average waiting time and queue length of the other traffic phases, QLTSTM compromises the green signal durations of one traffic phase for others. Fast reaction towards the changes of traffic flow input is also one of the capabilities of QLTSTM. The QLTSTM system's sensitivity towards the dynamic environment allows it to adapt in the dynamic changes of the incoming traffic flow as well as the vehicles in queue.

Q-learning algorithms' exploration in the dynamic traffic flows and exploit its best actions based on its experience has shown a good performance in the traffic signal timing plan management system. The traffic signal timing plan system is encouraged by the ability of Q-learning to learn and adapt with the dynamic changes of the traffic flow. Q-learning is assessed via the simulation to be a suitable method or technique to be implemented into the traffic flow control and optimization of urban traffic network system.

ACKNOWLEGDEMENT

The authors would like to acknowledge the financial assistance of the Ministry of Higher Education of Malaysia (MoHE) under Fundamental Research Grant Schemes (FRGS) 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).

REFERENCE

[1] F. Teklu, A. Sumalee, and D. Watling. "A Genetic Algorithm Approach for Optimizing Traffic Control Signals Considering Routing." *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, pp. 31-43, 2007.

[2] E. Azimirad, N. Pariz, and M.B.N. Sistani. "A Novel Fuzzy Model and Control of Single Intersection at Urban Traffic Network." *IEEE Systems Journal*, vol. 4, no. 1, pp. 107-111, March 2010.

[3] K. Khiang Tan, M. Khalid, and R. Yusof. "Intelligent Traffic Lights Control by Fuzzy Logic." *Malaysian Journal of Computer Science*, vol. 9, no. 2, pp. 29-35 1996.

[4] V. Gradinescu, C. Gorgorin, R. Diaconescu, and V. Cristea. "Adaptive Traffic Light using Car-to-Car Communication." *In Proceeding of Vehicular Technology Conference, 2007,* pp.21-25.

[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, no. 2, pp. 128-135, 2010.

[6] Z.Y. Liu, and F.W. Ma, "On-line Reinforcement Learning Control for Urban Traffic Signals." *In 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, no. 3, pp. 177-188, 2010.

[8] B.Abdulhai, R. Pringle, and G.J. Karakoulas. "Reinforcement Learning for True Adaptive Traffic Signal Control." *Journal of Transportation Engineering*, vol. 129, no.3, pp. 278-285, June 2003.

[9] C.J.C.H. Watkins, P. Dayan. "Technical Note: Q-learning." *Machine Learning*, vol. 8, no.3, pp. 279-292, May 1992.

[10] Y.K. Chin, L.K. Lee, N. Bolong, S.S. Yang, and K.T.K. Teo, "Exploring Q-Learning Optimization in Traffic Signal Timing Plan Management." *In Proceeding of 3rd International Conference on Computational Intelligence, Communication Systems and Networks*, 2011, pp. 269-274.

[11] Mi. Tokic, and G. Palm. "Value-Difference based Exploration: Adaptive Control between Epsilon-Greedy and Softmax." *In KI 2011: Advances in Artificial Intelligence*. J. Bach, S.Edelkamp, Ed. Berlin: Springer, 2011, pp. 335-346.

[12] N.J. Garber, and L.A. Hoel. *Traffic and Highway Engineering. 3rd Ed.* Pacific Grove, California: Thomson, 2002.