MULTI-OBJECTIVE REINFORCEMENT LEARNING FOR TRAFFIC SIGNAL COORDINATE CONTROL

YIN Shengchao, Department of Automation, Tsinghua University, yinsc07@mails.tsinghua.edu.cn

DUAN Houli, Department of Automation, Tsinghua University, duanhouli00@mails.tsinghua.edu.cn

LI Zhiheng, Department of Automation, Tsinghua University, zhhli@mail.tsinghua.edu.cn

ZHANG Yi, Department of Automation, Tsinghua University, zhyi@mail.tsinghua.edu.cn

ABSTRACT

In this paper, we propose a new multi-objective control algorithm based on reinforcement learning for urban traffic signal control, named multi-RL. A multi-agent structure is used to describe the traffic system where vehicles are regarded as agents. Reinforcement learning algorithm is adopted to predict the overall value of optimization objective given vehicles' states. The policy which minimizes the cumulative value of optimization objective is regarded as the optimal one. In order to make the method suitable to various traffic conditions, we also introduce a multi-objective control scheme in which optimization objectives are selected according to the real-time traffic state. The Optimization objectives include the number of vehicle stops, the average waiting time and maximum queue length of the next intersection. In addition, we also introduce the priority control of buses and emergent vehicles into our model. The simulation results indicate that our algorithm could perform more efficiently than traditional traffic light control methods.

Keyword: Traffic Signal Control, Multi-Agent System (MAS), Reinforcement Learning, Multi-objective Control

1. INTRODUCTION

Increasing traffic congestion on the road network makes the development of more intelligent

and efficient traffic control systems an urgent and important requirement. However, traffic system is a typical complex large-scale system mixed of a great number of interacting participants, which makes it very difficult to use traditional control algorithms to get satisfied control effect. Thus, intelligent algorithms have been used in attempts to build an efficient traffic control system, such as fuzzy control technology [1-2], artificial neural network [3-4] and genetic algorithm [5-6], which greatly improve the efficiency of traffic control.

Reinforcement learning is a category of machine learning algorithms including Q learning, temporal difference, SARSA algorithm and so on [7-9]. Reinforcement learning is to learn the optimal police by a trial-and-error process including perceiving states from the environment, choosing an action according to current states and receiving rewards from the environment. The policy which maximizes the expected long-term cumulative reward is considered as the optimal one. Reinforce learning is a self-learning algorithm which doesn't need an explicit model of the environment. Thus it can be applied in traffic signal control effectively to response to the frequent change of traffic flow and outperform traditional traffic control algorithm. Thorpe studied reinforcement learning for traffic light control in 1997. He used a neural network to predict the waiting time for all cars standing at the intersection and selected the best control policy using Sarsa algorithm [10]. Abdulhai et al. presented a basic framework of applying Q-learning to traffic signal control and got encouraging results while applying it to an isolated intersection [11]. MIKAMI el al. combined evolutionary algorithm and reinforcement learning for cooperative traffic signal control [12]. However, the above methods used traffic-light based value functions which means a large number of states need to be handled. Therefore, these methods suffer from the "dimension curse" and met with limited success when applied to large-scale road network. Wiering et al. utilized a car-based value function to solve this problem [13-14]. They made a predictor for each car to estimate the overall waiting time given possible choices of a traffic light using reinforcement learning, and selected the decision which minimized the sum of waiting time of all cars in the network. This method effectively reduced the states space and thus can be applied to large network control. Experiment in a network with 12 edge nodes and 16 junctions proved the effectiveness of this method.

However, Wiering's method used the overall waiting time as the optimization goal which is mainly suitable for the medium traffic condition. In practical traffic system, we should consider different optimization objectives in different traffic situation, which is called multi-objective control scheme in this paper. In the free traffic condition, we try to minimize the overall number of stops of vehicles in the network; while in the medium traffic condition, the overall waiting time is regarded as the optimal goal. In crowed traffic situation, queue spillovers must be avoided to keep the network from large-scale congestion, thus the queue length must be focused on [15]. Therefore, multi-objective control scheme can adapt to various traffic conditions and make a more intelligent control system. Therefore, we propose a multi-objective control strategy based on the Wiering's model. In our model, data exchange between vehicles and roadside equipments is necessary. Thus vehicular ad hoc network is utilized to build a wireless traffic information system.

This paper is organized as follows: in the second section, we will introduce how to describe the road network with a agent-based structure; section 3 describes how to exchange traffic data using the vehicle ad hoc network; in section 4, multi-agent traffic control using reinforcement learning is proposed; in section 5, the proposed method is applied to a road

network with eight intersections to prove its effectiveness; finally, in section 6, we draw the conclusion of this paper.

2. AGENT-BASED MODEL OF TRAFFIC SYSTEM

We use an agent-based model to describe the practical traffic system. Vehicles and traffic signal controllers in the road network are regarded as two types of agents. Data will be exchanged among these agents. A typical road network is built based on the Wiering's model [14] as shown in figure 1. There are six possible settings for each traffic controllers to prevent accidents: two traffic lights from opposing directions allow cars to go straight ahead or to turn right (2 possibilities), two traffic lights at the same direction of the intersection allow the cars from there to go straight ahead, turn right or turn left (4 possibilities). Road lanes are discretized into a number of cells at each different traffic light. The capacity of each road lane is defined according to its practical length. At each time step, new cars are generated with a particular destination and enter the network from outside. After new cars have been added, traffic light decisions are made and each car moves to the subsequent cell if it is not occupied or the car's predecessor is moved forward. Thus, each car is at a specific traffic node (**node**), a direction at the node (**dir**), a position in the queue (**place**) and has a particular destination (**des)**. Thus we can use [node, dir, place, des] ([n, d, p, des] for short) to denote the state of each vehicle [13]. As mentioned before, a multi-objective control scheme is adopted in this method. The optimization objectives include waiting time, stops and queue length, which will be selected according to the traffic condition. We use $Q([n,d,p,des],action)$ to denote the total expected value of optimized indices for all traffic lights for each car until it arrives at the destination given its current node, direction, place and the decision of the light. The optimal action of node j is determined by the following formulation:

$$A_j^{opt} = \arg\max_{A_j} \sum_{i \in A_j} \sum_{(n,d,p,des) \in queue_i} Q([n,d,p,des],red) - Q([n,d,p,des],green) \qquad (1)$$

It should be noticed $Q([n,d,p,des],action)$ doesn't only refer to the waiting time but also stops and queue lengths. This is the most import difference between our model and Wiering's model, which will be introduced in detail in section 4.

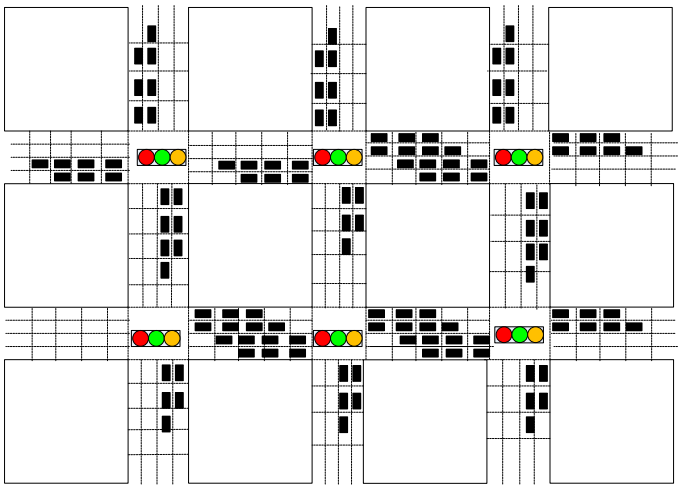

Figure 1 – Agent-based traffic model illustration

3. TRAFFIC INFORMATION EXCHANGE SYSTEM

We need to exchange a lot of information during the control process. Thus, a wireless traffic information exchange system based on vehicular Ad hoc network is built to exchange data between vehicles and signal controllers. The illustration of such an information exchange system is showed in figure 2. Vehicles in the network all have the ability of communicating with each other and the controllers. Thus necessary information can be collected through the intercommunication of vehicles and controllers. The data to be collected includes:

- Traffic flow through each intersection in each time step;
- Queue length at each traffic light in each time step;
- Type of each vehicle (car, bus or emergent vehicle);
- Destination of each vehicle;
- Node where each vehicle stands at;
- Direction each vehicle moving towards;
- Position in the queue where each vehicle stands at;
- Total waiting time each vehicle used to pass through the network;
- Total number of stops each vehicle used to pass through the network.

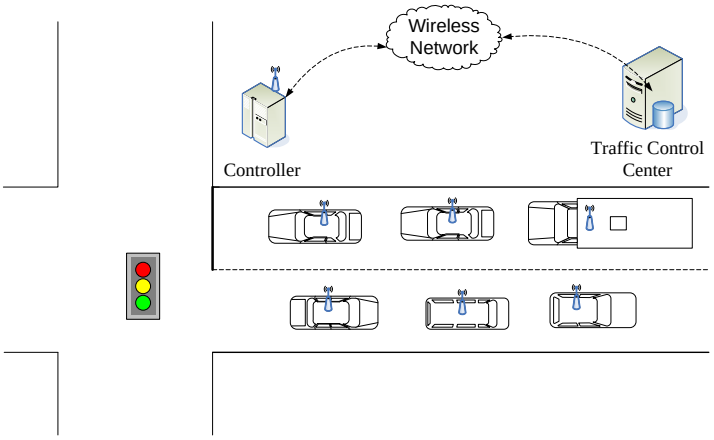

Figure 2 – Illustration of Traffic Information Exchange System

4. MULTI-OBJECTIVE CONTROL ALGORITHM BASED ON REINFORCEMENT LEARNING (MULTI-RL ALGORITHM)

We extend the control algorithm to a multi-objective scheme by selecting optimization objective according to real-time traffic condition. In addition, we think some special vehicles such as buses and ambulances need a priority control and thus should be considered separately.

The multi-objective control algorithm considers three types of traffic conditions as follows. The method to estimate traffic conditions should be defined carefully according to the actual situation of road network.

1) Free traffic condition

Under this condition, we aim to minimize the number of stops, in other words, we expect to make the vehicles pass through the network with the fewest stops. Thus, the cumulative number of stops is selected as the optimization objective.

The number of stops will increase when a vehicle moving at a green light in current time step meet a red light in the next time step. Therefore, we denote $Q([node, dir, pos, des], \text{Green})$ as the expected cumulative number of stops since this number will not change when the traffic light is red in current time step. The iterative formulation of $Q([node, dir, pos, des], \text{Green})$ is shown as follows.

$$Q([node, dir, pos, des], \text{Green}) = \sum_{(node', dir', pos')} P(\text{Red} \mid [node', dir', pos', des])$$
$$(R([node, dir, pos, des], [node', dir', pos', des]) + \gamma Q([node', dir', pos', des'], \text{Green})) \quad (2)$$

Where $[node', dir', pos', des]$ means the state of a vehicle in next time step; $P(\text{Red} \mid [node', dir', pos', des])$ gives the probability that the traffic light turns red in next time step; $R([node, dir, pos, des], [node', dir', pos', des])$ is a reward function as follows: if a car stays at the same traffic light, then $R = 1$, otherwise, $R = 0$ (the car gets through this intersection and enters the next one); γ is the discount factor($0 < \gamma < 1$) which ensure the Q-values are bounded. The probability that a traffic light turns red is calculated as follows.

$$P(\text{Red} \mid [node, dir, pos, des]) = \frac{C([node, dir, pos, des], \text{Red})}{C([node, dir, pos, des])} \quad (3)$$

Where $C([node, dir, pos, des])$ is the number of times a car in the state of $[node, dir, pos, des]$, $C([node, dir, pos, des], \text{Red})$ is the number of times the light turns red in such state.

2) Medium traffic condition

Under medium traffic condition, we focus on the overall waiting time of vehicles, which is the same as in Wiering's model [13-14]. $Q([node, dir, pos, des], \text{action})$ is used to denote the total waiting time for all traffic lights for each car until it arrives at the destination given its current state and the action or the light. $V([node, dir, pos, des])$ denotes the average waiting time (without knowing the traffic light decision) for a car at $node, dir, pos$ until it reaches its destination. $Q([node, dir, pos, des], \text{action})$ and $V([node, dir, pos, des])$ are iteratively updated as follows.

$$V([node, dir, pos, des]) = \sum_{L} P(L \mid [node, dir, pos, des]) Q([node, dir, pos, des], L) \quad (4)$$

$$Q([node, dir, pos, des], L) = \sum_{(node', dir', pos')} P([node, dir, pos, des], L, [node', dir', pos', des])$$
$$(R([node, dir, pos, des], [node', dir', pos', des]) + \gamma V([node', dir', pos', des'])) \quad (5)$$

Where L is the traffic light state (red or green), $P(L \mid [node, dir, pos, des])$ is calculated in the same way as equation 3, $R([node, dir, pos, des], [node', dir', pos', des])$ is defined as follows: if a car stays at the same place, then $R = 1$, otherwise, $R = 0$ (the car can move forward).

3) Congested traffic condition

Under congested traffic condition, we must do out best to avoid the queue spillovers, which will degrade the traffic control effect and probably cause large-scale traffic congestion [15]. Therefore, the queue length is taken into consideration when we design the Q learning procedure. Denote the maximum queue length at the next traffic light tl' as $K_{tl'}$, shortly written as K. When traffic light is red, no vehicle will pass through to the next light. Thus the equations at red light doesn't change, we focus on the function when light is green. Then equation 5 can be rewritten as follows.

$$Q([node, dir, pos, des], \text{Green}) = \sum_{(node', dir', pos')} P([node, dir, pos, des], \text{Green}, [node', dir', pos', des])$$
$$(R([node, dir, pos, des], [node', dir', pos', des]) + \alpha R'([node, dir, pos, des], [node', dir', pos', des])$$
$$+\gamma V([node', dir', pos', des'])) \qquad (6)$$

$$Q([node, dir, pos, des], \text{Red}) = \sum_{(node', dir', pos')} P([node, dir, pos, des], \text{Red}, [node', dir', pos', des])$$
$$(R([node, dir, pos, des], [node', dir', pos', des]) + \gamma V([node', dir', pos', des'])) \qquad (7)$$

Where $Q([node, dir, pos, des], L)$ and $V([node, dir, pos, des])$ have the same meanings as under medium traffic condition. Compared equation 6 with equation 5, another reward function $R'([node, dir, pos, des], [node', dir', pos', des])$ is added to indicate the influence from traffic condition at the next light. $R([node, dir, pos, des], [node', dir', pos', des])$ is the reward of vehicles' waiting time while $R'([node, dir, pos, des], [node', dir', pos', des])$ indicates the reward from the change of the queue length at the next traffic light. α is an adjusting factor.
$R([node, dir, pos, des], [node', dir', pos', des])$ is defined as follows: if a car stays at the same place, then $R = 1$, otherwise, $R = 0$ (the car can move forward).
$R'([node, dir, pos, des], [node', dir', pos', des])$ is defined as follows: if a car passes through the current intersection to the next traffic light, which means the queue length at the next traffic light will increase by 1 in a short time, then $R = 1$, otherwise, $R = 0$.
Given the capacity of the lane of next traffic light is L, then the adjusting factor α is determined by the queue length $K_{tl'}$ as follows.

$$\alpha = \begin{cases} 0 & if\ K_{tl'} \le 0.8L \\ 10(\frac{K_{tl'}}{L} - 0.8) & if\ 0.8L < K_{tl'} \le L \\ 2 & if\ K_{tl'} > L \end{cases} \qquad (8)$$

Through the definition we can find that α will increase sharply when the queue length approaches the capacity of the lane, which means queue spillovers would like to happen. Thus, under such a situation, $Q([node, dir, pos, des], \text{Green})$ will increase sharply and make the gain of this police decrease. Therefore, the green phase length and the number of vehicles allowed to pass through will be decreased until the queue at the next light has been dispersed. The largest value of α is set to 2 in this paper, but you can adjust its value according to the practical traffic condition.

4) Priority control for buses and emergent vehicles

When buses or emergent vehicles (fire trucks or ambulances) enter the road network, they should have priority to pass through. It is necessary to realize the priority control of these special vehicles with least disturbance to the regular traffic order. Thus, we revise the equation 5 as follows. A priority factor β is added to describe the emergent degree of these special vehicles, which need be determined uniformly by the traffic management department.

$$Q([node, dir, pos, des], L) = \sum_{(node', dir', pos')} P([node, dir, pos, des], L, [node', dir', pos', des])$$
$$(\beta R([node, dir, pos, des], [node', dir', pos', des]) + \gamma V([node', dir', pos', des'])) \quad (9)$$

5. CASE STUDIES

We have done some case studies to prove the effectiveness of our model. Since it is very hard to apply a signal control model to real traffic system management, traffic simulation was chosen to do the case studies. Paramics V6.3 was selected as the simulation platform because it is a professional traffic simulation tool which is recognized by traffic engineers all over the world. A practical road network within Beijing Second Ring Road was modeled in Paramics as shown in figure 3. This is a network with 7 intersections (N1-N7) and 8 OD zones (Zone1-Zone8). Intersections N1-N7 corresponds to the real intersections Xiaoweihutong, Dongdansantiao, Jingyuhutong, Dengshidongkou, Dengshikou, Wangfujingbeikou and Taiwanfandian.

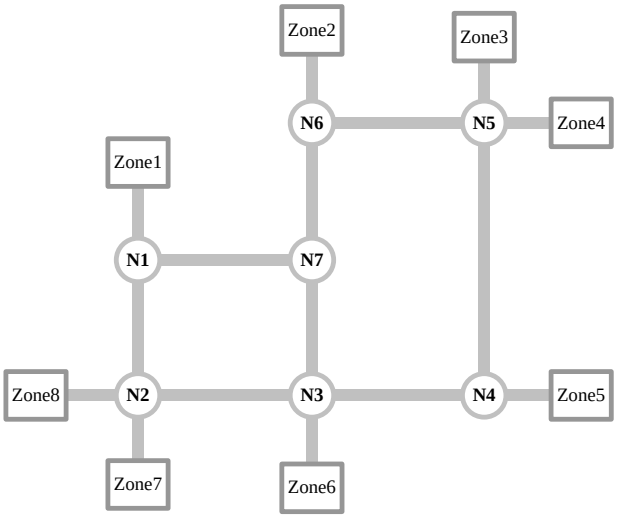

Figure 3 –Sketch diagram of a practical road network in Beijing

The simulation ran for 10000 time steps, the former 4000 steps was the learning process, and the latter 6000 steps was used to collected the simulation results. Factor γ is set to be 0.9 and β is set to be 3. The lanes in the network are divided into cells with length of 7.5 m. The capacity of the lanes equals to the number of the cells.

We compared our method with fixed control, actuated control and also Wiering's method. The set of fixed control is: the cycle is 2 minutes and the green time is equally assigned to each phase. In the actuated control strategy, the minimum green time is 10s, the maximum green

time is 50s, the extension of green time is set to 4s. Parameters of Wiering's method are the same as our model under the medium traffic condition.

We wanted to estimate of effectiveness of the multi-objective scheme, thus we estimate the control effects of these four algorithms under different traffic conditions. So we changed the traffic volume entering the network from 30 to 270, and estimated the average waiting time, number of stops and maximum queue length of these four methods.

In our model, when the traffic volume entering the network in a minute is less than 90, it is regarded as free traffic; when the volume is larger than 90 but less than 180, it is regarded as medium traffic; when the traffic volume is larger than 180, it is regarded as congested traffic condition.

1) Comparison of average waiting time

The comparison of average waiting time with respect to the increasing of traffic volume is shown in figure 4. Fixed means the fixed control strategy, actuated means the vehicle actuated method, RL means the algorithm proposed by Wiering [13-14] and multi-RL means the model proposed in this paper.

It is obvious to see when the traffic volume is less than 90, which means the traffic state is free, number of stops under the multi-RL control will be less than those under other control strategies. This is because the multi-RL is the only one who aims to minimize the number of stops. However, with the increase of traffic volume, the multi-RL method changes its objective, and the actuated control gets the minimum stops.

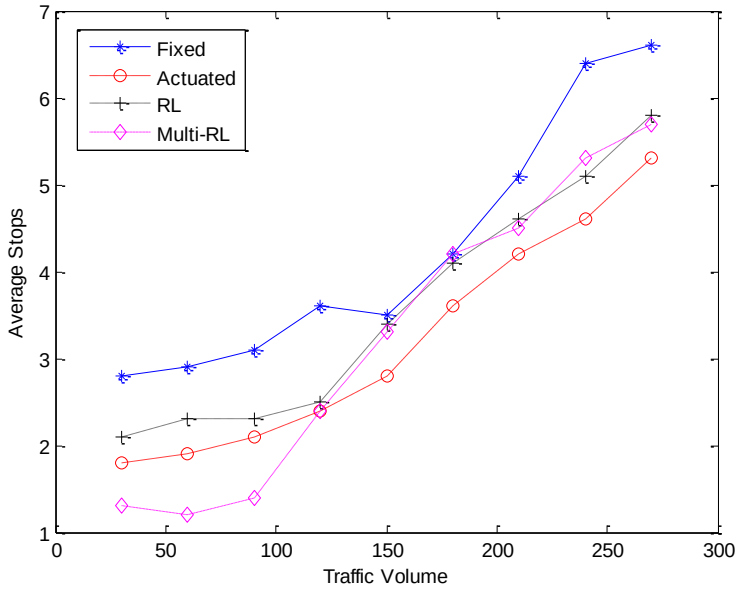

Figure 5 –Control effects comparison estimated by average stops

2) Comparison of number of stops

The comparison of average waiting time with respect to the increasing of traffic volume is shown in figure 5. Since the multi-RL is the same as RL method under the medium traffic

condition, they have the same average waiting time in the middle. Under the free traffic state, RL gets the minimum waiting time because this is its optimization objective. It should be noticed multi-RL gets the minimum waiting time when the traffic is congested. That indicates although the RL aims to minimize waiting time, the queue spillover which is not considered will decrease the traffic efficiency and increase the waiting time.

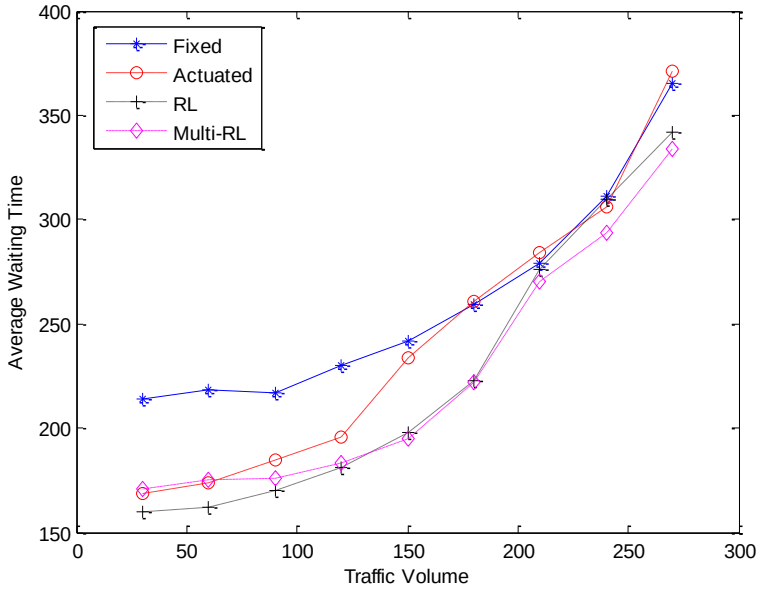

Figure 5 – Control effects comparison estimated by average waiting time

3) Comparison of maximum queue length

The comparison of average waiting time with respect to the increasing of traffic volume is shown in figure 6.

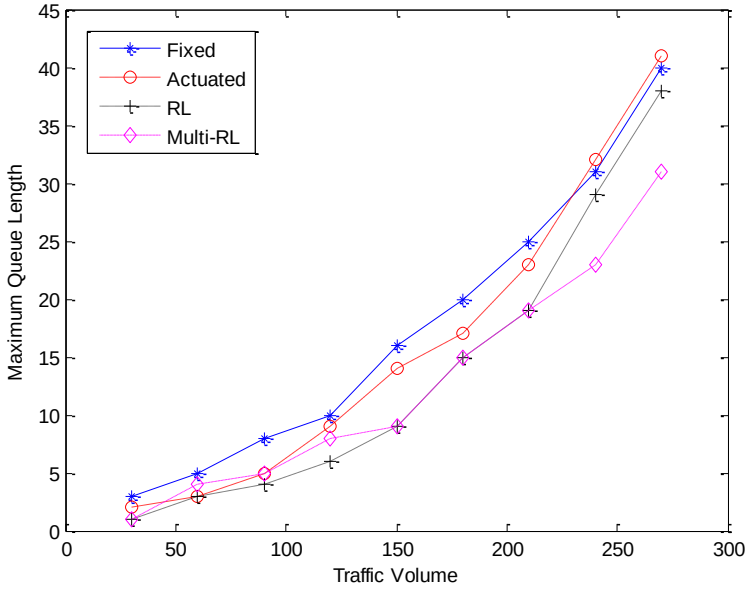

Figure 6 – Control effects comparison estimated by maximum queue length

The maximum queue length exceeds 40 under fixed control, which indicates there must be some queue spillovers. This is taken into consideration in the multi-RL thus we get a short queue under congested traffic condition.

6. CONCLUSION

In this paper, we have presented the multi-objective control algorithm based on reinforcement learning. The simulation indicated that: the multi-RL got the minimum stops under free traffic, although not the minimum waiting time; the multi-RL had the similar performance with the RL method under medium traffic, which was better than fixed control and actuated control; under congested condition, multi-RL could effectively prevent the queue spillovers to avoid large scale traffic jams. It also should be noticed multi-RL is a car-based algorithm, which is less time-consuming than the light-based reinforcement learning algorithms [13].
However, there are still some system parameters that should carefully be determined by hand. For, example, the adjusting factor α indicating the influence of the queue at the next traffic light to the waiting time of vehicles at current light under congested traffic condition. This is a very important parameter, which we should further research its determining way based on traffic flow theory. In addition, some phenomenon in real traffic system such as the lane changing of cars will influence their travel time. We should further take these into consideration and build a model more close to the real traffic system.

7. Acknowledgments

This work is supported by the National High Technology Research and Development Program ("863" Program) of China, contract number 2007AA11Z215; by the Key Project of Chinese National Programs for Fundamental Research and Development (973 program), contract number 2006CB705506; by Chinese National Natural Science Foundation, contract number 60834001, 60774034, 60721003, 50708055.

Reference

C. P. Pappis and E. H. Mamdani (1977). A Fuzzy Logic Controller for a Traffic Junction. IEEE Transactions on Systems, Man and Cybernetics, 7, 707-717.

Mohamed B. Trabia, et al (1999). A Two-stage Fuzzy Logic Controller for Traffic Signals. Transportation Research Part C, 7, 353-367.

J. C. Spall, and D. C. Chin (1997). Traffic-Responsive Signal Timing for System-wide Traffic Control. Transportation Research Part C: Emerging Technologies, 5(3), 153-163.

Liu Zhiyong, et al (1997). Hierarchical Fuzzy Neural Network Control for Large Scale Urban Traffic Systems. Information and Control, 26(6), 441-448.

M. D. Foy, et al (1992). Signal Timing Determination Using Genetic Algorithms. Transportation Research Record 1365, National Research Council, Washington, D.C., 108-115.

B. Park, et al (2000). Enhanced Genetic Algorithm for Signal Timing Optimization of

Oversaturated Intersections. Transportation Research Record 1727, National Research Council, Washington, D.C., 32-41.

R. Sutton (1988). Learning to Predict by the Methods of Temporal Difference. Machine Learning, 3, 9-44.

Watkins C (1989). Learning from Delayed Rewards, PHD thesis, King's college, UK.

L.P. Kaelbling, et al (1996). Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research, 4, 237-285.

Thorpe T. (1997). Vehicle Traffic Light Control Using SARSA, Master thesis, Colorado State University, USA.

B. Abdulhai, et al (2003). Reinforcement Learning for True Adaptive Traffic Signal Control. ASCE Journal of Transportation Engineering, 129(3), 278-285.

Miami S, Kakazu Y(1994). Genetic Reinforcement Learning for Cooperative Traffic Signal Control. Evolutionary Computation, 1, 223-228.

M. Wiering, et al (2004). Intelligent Traffic Light Control. Technical Report UU-CS-2004-029, University Utrecht.

M. Wiering(2000). Multi-Agent Reinforcement Learning for Traffic Light Control. Machine Learning: Proceedings of the 17th International Conference (ICML' 2000), 1151-1158.

C. F. Daganzo (1998). Queue Spillovers in Transportation Networks with a Route Choice. Transportation Science, 32(1), 3-11.