KOSTANDINA VELJANOVSKA, Ph.D.
E-mail: kostandina@rocketmail.com,
Sv. Kliment Ohridski University,
Faculty of Administration and Information
System Management
Partizanska bb, 7000 Bitola, Republic of Macedonia
KRISTI M. BOMBOL, Ph.D.
E-mail: kristi.bombol@uklo.edu.mk
Sv. Kliment Ohridski University,
Faculty of Technical Sciences
POB 99, 7000 Bitola, Republic of Macedonia
TOMAŽ MAHER, Ph.D.
E-mail: tmaher@fgg.uni-lj.si
University of Ljubljana, Faculty of Civil and
Geodetic Engineering
Jamova 2, SI-1000 Ljubljana, Republic of Slovenia

Intelligent Transport Systems (ITS)
Preliminary Communication
Accepted: May 12, 2009
Approved: Mar. 19, 2010

REINFORCEMENT LEARNING TECHNIQUE IN MULTIPLE MOTORWAY ACCESS CONTROL STRATEGY DESIGN

ABSTRACT

An appropriately designed motorway access control can decrease the total travel time spent in the system up to 30% and consequently increase the merging operations safety. To date, implemented traffic responsive motorway access control systems have been of local or regulatory type and not truly adaptive in the real sense of the meaning. Hence, traffic flow can be influenced positively by numerous intelligent transportation system (ITS) techniques. In this paper a contemporary approach is presented. It considers the design philosophy of an optimal and adaptive closed-loop multiple motorway access control strategy. The methodology proposed uses the artificial intelligence technique - known as reinforcement learning (RL) with multiple agents, and applies the Q-learning algorithm. One segment of the motorway network with three lanes in each direction and three motorway entries was designed. The detectors and traffic signals were placed at the entries (ramps). Traffic flows and traffic occupancy on the main line as well as the traffic demand on the motorway entries were taken as input model variables. The output variables referred to the travel speed on the corridor, the total travel time, and the total stop time. VISSIM micro-simulator and direct programming of the simulator functions were used in order to implement the RL technique. The peak hour was chosen for the time of simulation.

The model was tested in two phases. Its effectiveness was compared to ALINEA. It was observed that the proposed strategy was capable of responding both to dynamic sensory inputs from the environment and to dynamically changing environment. The model of the environment and supervision were not required. The control policy changed as response to the inherent system characteristic changes. It was confirmed that the strategy was truly adaptive and real-time responsive to the traffic demand on the corridor.

KEYWORDS

motorway access, traffic flows, control, strategy, artificial intelligence, Q-Learning, simulation

1. INTRODUCTION

Recurrent and non-recurrent motorway congestion leads to delays, reduced traffic safety, increased fuel consumption, and serious air pollution as well. Such congestion limits the motorway throughput at times when it is most necessary, i.e. during the peak hour. The throughput becomes even more critical when non-recurrent congestion occurs. Building new motorways will leave current motorway infrastructure insufficiently utilized. On the contrary, traffic flows can be positively influenced by numerous intelligent transportation system (ITS) techniques.

The examples of motorway access control systems are numerous. ALINEA [5, 11, 17, 18] is the first control strategy on a local level and is based on direct implementation of classical control theory with feedback. Other efforts include genetic fuzzy approach, artificial neural networks, and two-level motorway access control approach [9, 21, 22].

All the existing motorway access control algorithms, although traffic responsive, are not truly adaptive to traffic parameter changes [19, 20, 14]. Most of them are of local regulatory type [4, 5]. Adaptive in this sense is opposed to the common controversial interpretation of the term in literature. It means more than giving a real time traffic response only. Additionally, the control policy changes itself as a response to the inherent systems characteristics. In other words, in order to be truly adaptive, the system should be capable of learning continuously [4].

In this respect, by implementing the information technology methodology, i.e. the specific artificial intelligence technique, a truly adaptive strategy for multiple motorway access control can be designed and developed. The main research hypothesis refers to the statement that motorway access control can be a completely adaptive and optimal closed loop control strategy that minimizes total travel time on the corridor.

This paper is an attempt to go a step further and use the adaptive control strategy when the level of traffic density necessary to be maintained is not predefined – a situation wherein the strategy itself learns how to minimize the total travel time spent in the system. Furthermore, the agents continuously learn by themselves and adapt to the environment changes accordingly.

2. ARTIFICIAL INTELLIGENCE TECHNIQUE USED

Reinforcement learning (RL) is a machine learning technique which does not require supervised training as it is the case with other learning techniques such as neural networks. It is based on goal-directed learning from interaction with an environment, i.e. *what to do* or *how to map* situations or states towards actions in order to maximize a numerical reward signal. By trying, exploring, and exploiting actions in an iterative process, the learner – the so-called autonomous agent, senses and learns in its environment how to choose the optimal action, or the actions that yield the cumulative reward.

More specifically, the agent and the environment interact at each sequence of discrete time steps $t = 0, 1, 2, 3 \ldots$. At each time step, the agent, *t*, receives some representation of the environment's state, here expressed as $s_t \in S$ (where S is the set of possible states), and accordingly selects an *action*, here expressed as $a_t \in A(s_t)$ (where $A(s_t)$ is the set of actions available in state s_t). One time step later, partly - as a consequence of its action, the agent receives a numerical *reward*, $r_{t+1} \in R$, and finds itself in a new state, S_{t+1}. A trainer may provide a reward or penalty to indicate the desirability of the resulting state. The transition from state to state is expressed as

$$S_0 \xrightarrow[r_0]{a_0} S_1 \xrightarrow[r_1]{a_1} S_2 \xrightarrow[r_2]{a_2} \tag{1}$$

At each time step, the agent implements mapping of the state representations and the probabilities of selecting each possible action. This mapping is called the agent's *policy*. The most important features of the agent are trial and error search and delayed reward.

In RL, the agent goal is formalized in terms of a special signal called a *reward* that passes from the environment to the agent. The agent tries to select actions so that the sum of the discounted rewards it receives gets maximized, here expressed as

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \tag{2}$$

where R_t is the expected discounted reward, r_t is the reward in the time step *t,* and γ is the *discount rate*.

In particular, it chooses at to maximize the expected *discounted return*, where γ is a parameter between zero and one.

Almost all reinforcement learning algorithms are based on *estimating value functions i.e.*-functions of states (or state - action pairs) that estimate *how good* it is for the agent to be in a given state. This is explained in 2.1.

2.1 Q-Learning

One of the most important improvements in RL was the development of an off-policy Temporal Difference (TD) control algorithm known as *Q-learning*. This algorithm, developed by Watkins, has been researched most frequently, both theoretically and practically. This is mainly due to its origination from the concept and principles of Dynamic Programming (DP) [1]. Thus related to DP, Q-learning integrates planning and learning unlike other reinforcement algorithms [2]. One of the most important features of this algorithm is that it does not require a pre-specified model of the environment upon which to base its action selection. Instead, only relationships between states, actions, and rewards are learned. Almost all of the traffic control methods, except the recent ones, usually require prespecified models of traffic flow to generate short-term predictions of traffic conditions or to assess the impacts of possible control decisions [3].

The Q-learning task can be defined as acquiring optimal policy π by learning value function V^* of the optimal policy π^*, provided by perfect knowledge of the immediate reward function *r* and the state transition function δ. When the agent knows the functions *r* and δ used by the environment to respond to its actions, then it can calculate optimal action for any state *s* as

$$\pi^*(s) = \arg\max_{\alpha} [r(s,a) + \gamma V^*(\delta(s,a))]. \tag{3}$$

If the evaluation function $Q(s,a)$ represents the reward, which is received for executing action a from state s and to which the value discounted by γ is added, here expressed by

$$Q(s,a) = r(s,a) + \gamma V^*(\delta(s,a)), \tag{4}$$

then the agent will select optimal actions even when it has no knowledge of the functions *r* and δ, that is to say

$$\pi^*(s) = \arg\max_{\alpha} Q(s,a) \tag{5}$$

In this case, independently of the policy being followed, the learned action-value function Q directly approximates Q^*, that is to say the optimal action-value function.

It is assumed that under certain conditions in a deterministic world (for MDP) estimated value for $Q^\wedge$ will converge to true Q value. Different authors have made some modifications of the original algorithm introducing learning rate α expressed by

$$Q(s,a) \longleftarrow Q(s,a) + \alpha[r + \gamma \max_{\alpha'} Q(s',a') - Q(s,a)], \tag{6}$$

where $Q(s,a)$ is the function of the action reward, α is the learning rate ($0 < \alpha < 1$), γ is the decrease rate parameter, $Q(s',a')$ is the function of the new action value a' for the new state s'.

Learning rule used in this research is defined by Q-learning algorithm by Watkins for non-deterministic processes [16]. This is the case because the probability distributions both for the reward function $r(s,a)$ and for the transition function $\delta(s,a)$ depend on s and a only. They do not depend on previous states or actions as it is a non-deterministic Markov decision process (MDP). Since traffic is a stochastic process, in the learning rule

$$Q(s,a) = r(s,a) + \gamma V^*(\delta(s,a)) \tag{7}$$

the non-deterministic environment has to be accommodated. The function of the action reward $Q(s,a)$ is redefined as a value expected from the previously defined value for deterministic case. Hereby, the rule becomes

$$\begin{aligned} \hat{Q}_n(s,a) \longleftarrow & (1-\alpha_n)\hat{Q}_{n\text{-}1}(s,a) + \\ & + \alpha_n\left[r + \max_{\alpha'} \hat{Q}_{n\text{-}1}(s',a')\right]. \end{aligned} \tag{8}$$

In equation (8), $\hat{Q}_n(s,a)$ is a value expected from the previously defined value for deterministic case of the action function a for state s, α_n is the learning rate, $\hat{Q}_{n\text{-}1}(s',a')$ is the value expected from the previously defined value of the new action a' for the new state s'.

The learning rate α_n is expressed by

$$\alpha_n = \frac{1}{1 + visits_n(s,a)} \tag{9}$$

In the above equation s and a are the state and action updated during the n-th iteration, and $visits_n(s,a)$ is the total number of times that this state-action pair has been visited up to including the n-th iteration. This rule is suitable for deterministic case when α_n is 1. As n increases α_n decreases. By reducing α_n at an appropriate rate during training, convergence of Q values can be achieved. In order to speed up the learning process, fixed α_n was used in our experiments.

3. MODEL TESTING

In order to test the control strategy, a few scenarios were divided into two test cases in accordance with the traffic parameters:

- the first test case - coordinated control and parameters measurements taken at the motorway exit, with known traffic demand on the main line (*Figure 1*);
- the second test case - coordinated control and measurements taken downstream at each motorway entry, with unknown traffic demand on the main line (*Figure 2*). During this test case two types of scenarios were developed: 1 - testing when there is no traffic congestion, 2 - testing when there is traffic congestion in the corridor.

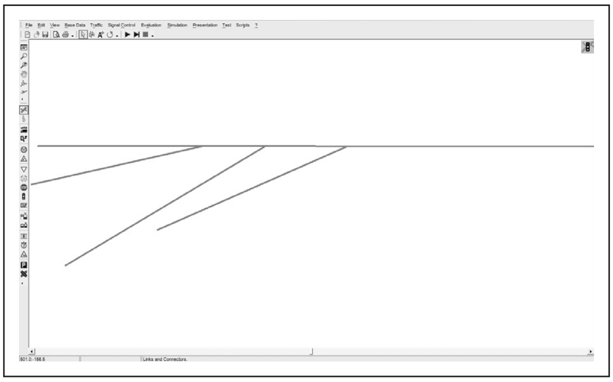
File Edit View Base Data Traffic Signal Control Evaluation Simulation Presentation Test Scripts ?
601.0:-166.6
Links and Connectors.

Figure 1 - First test case layout

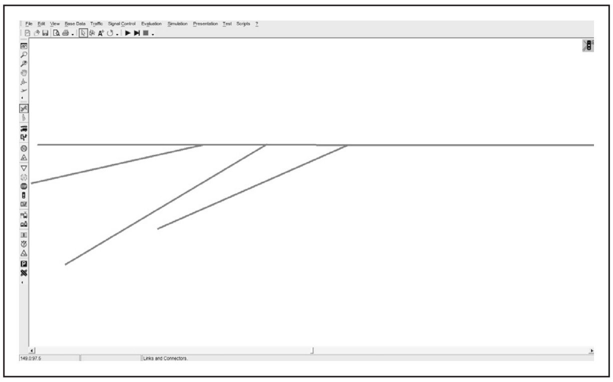
File Edit View Base Data Traffic Signal Control Evaluation Simulation Presentation Test Scripts ?
149.0:97.5
Links and Connectors.

Figure 2 - Second test case layout

In order to estimate the feasibility of the suggested strategy for optimal adaptive coordinated control of the motorway entry ramp, the results from the agents that learn were compared to the results from the case with no control strategy and to those from the case with ALINEA control - the widely implemented control strategy used as a regulator.

The results gained from the simulations with no control strategy were taken as the base case and the rest of the results that were compared to it were estimated. Testing was conducted after sufficient number of iterations with different numbers of states and after Q-values convergence [4].

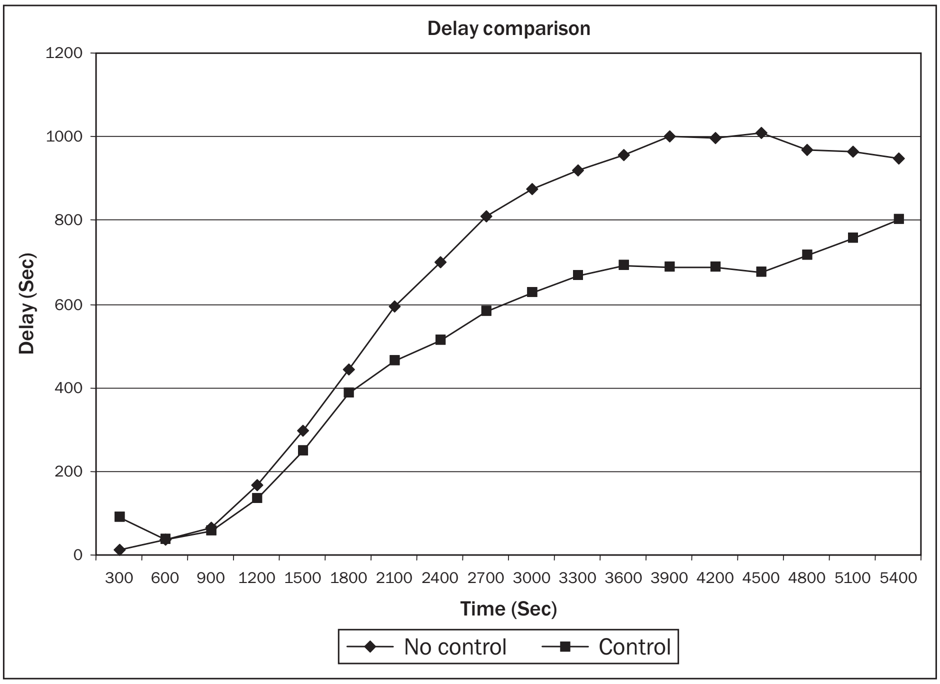

Figure 3 - Delay comparison

The above presented strategy for optimal adaptive coordinated motorway access control uses the so-called look-up table. [4]

4. DISCUSSION

Within the *first test case* (coordinated motorway access control, measurements at the exit of the corridor, traffic demand known), improvements were as follows:

- savings in travel time up to 14.50%;
- delay decrease by 26%;
- average stop time per vehicle decrease by 37%;
- average number of stops per vehicle decrease by 35%, and
- the number of vehicles exiting the network increase by 14%.

It is evident that this type of control strategy needs a longer phase of learning for the agents, which makes the strategy not efficient enough. Therefore, localized motorway entry access was implemented, whereas traffic parameters were measured on the mainline downstream of each access (*the second test case*). During this test case two types of testing (scenarios) were performed:

1. testing with *no traffic congestion* present;
2. testing with *traffic congestion* present.

After performing tests with data showing *no traffic congestion present* (*Scenario 1*), it was noticeable that there were significant improvements regarding:

- delay (decreased by 30%) (*Figure 3*);
- average stop time per vehicle (reduced by 78%);
- average number of stops per vehicle (reduced by 80%) proving the smoothness of traffic flow;
- longer traveling, evident travel time and delay decrease and a significant difference after one hour of travel.

There was very little improvement in:

- travel time (reduced by 3.29%);
- number of vehicles exiting the corridor (increased by 3%);
- speed change (increased by 0.33% only).

It was noticeable that the strategy followed real-time traffic parameters change, particularly during the transition from the state of congestion to the normal state. The results from implementation of ALINEA for the same effectiveness parameters were similar to the corresponding results gained by the suggested control strategy. This similarity could be explained with the fact that there was no recurrent congestion on the corridor, which made this strategy inferior as compared to ALINEA.

Regarding travel time savings, speed increase, and the number of vehicles exiting the corridor, the results gained with ALINEA were not very promising. This is important because the ALINEA strategy implementation requires some parameter calibrations to be made for the particular motorway and for the corresponding traffic demand. However, the above coordinated control strategy testing can be performed on unknown traffic demand. Therefore, in the case with *no traffic congestion*, the suggested strategy could be implemented with learning performed with traffic demand similar to the one preceding the implementation.

During the *second test case* (with traffic congestion on the corridor and with unknown traffic demand) the Q-learning strategy shows extraordinarily good re-

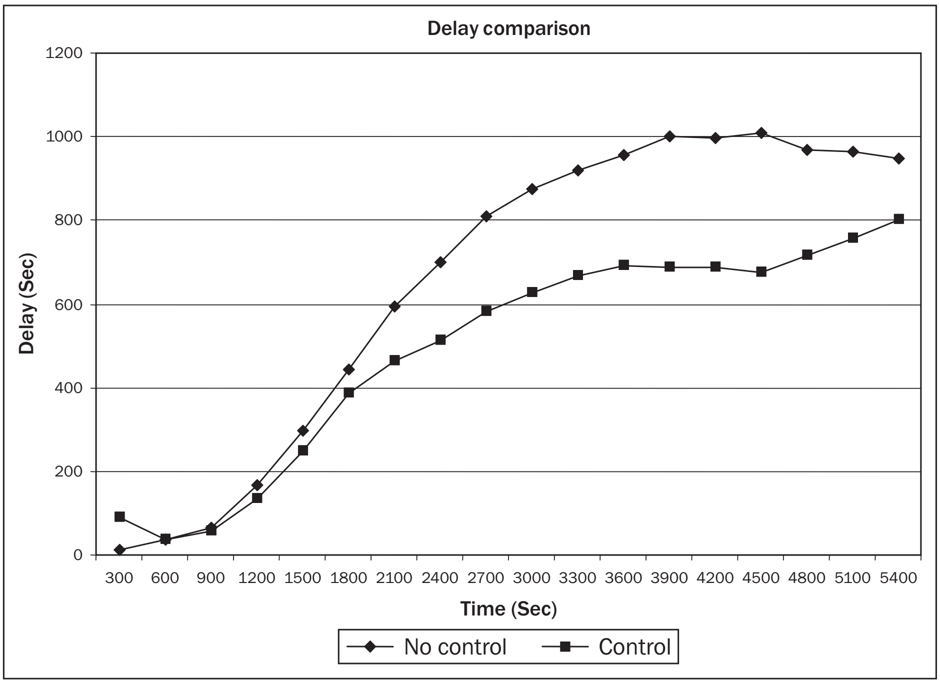

Figure 4 - Delay comparison

sults after relatively small number of iterations (about 1500). The outcome results were as follows:

- savings in travel time increase by 15%;
- delay decreases by 26% (*Figure 4*);
- average stop time per vehicle decreases by 38%;
- average number of stops per vehicle decreases by 35%;
- increase in the number of vehicles exiting the network by 10%;
- speed increase by 9.85%.

Improvements were almost doubled compared to the results with ALINEA implementation with the same measures of effectiveness (8.41%, 13%, 20%, 19%, 6.22%, and 3.55%, respectively). It was obvious that the strategy adjusted itself to the traffic conditions, i.e. it is adaptive and responds to the real-time traffic demand. Thus, the main research hypothesis stated at the very beginning has been proven [4].

The best improvement was achieved in the case of control implementation with data showing *no congestion* (for the average stop time per vehicle and average number of stops per vehicle).

Regarding all the measures of effectiveness, the best results were gained when control strategy was implemented *on unknown traffic demand with congestion*. This shows that the suggested strategy is feasible for coordinated motorway access control that is optimal, adaptive, and traffic responsive.

After the testing with data where *there is traffic congestion* and *unknown traffic demand* on the corridor, the strategy that uses Q-learning showed extraordinarily good results after relatively small number of iterations. Thus, its feasibility and efficiency have been confirmed as well.

Suggested coordinated control strategy proves better than ALINEA in relation to the average stop time per vehicle and average number of stops per vehicle during the peak hour. The evidence of this lies in the smoothness of the traffic flow with no interruptions in terms of *stop-and-go*. This leads to reduced fuel consumption per vehicle, reduced air pollution, and reduced environmental pollution as well. [4]

5. CONCLUSION

Bearing in mind the results of the model testing, it can be concluded that an optimal adaptive coordinated motorway access control is feasible for performing multiple motorway access control.

This research opens broad possibilities for reinforcement learning technique implementation in traffic control. Some of the steps in scientific research to follow are to deal with coordinated control for non-congested traffic, traffic signal control on isolated intersections, and examination of the model efficiency after implementation.

This research shows the implementation of real-time traffic control strategy. Several facts confirm its uniqueness such as:

1. the strategy requires no environment modeling;
2. the strategy is truly adaptive;
3. supervision is not necessary,
4. no need for traffic parameters prediction;

5. the best optimal control strategy, based on the current traffic state only and on the current control conditions, simplifies the approach;
6. the strategy can be implemented in real time since the model requires neither simulation steps to be performed nor any calculations to be made during the implementation phase

Taking the above into account, the conclusion follows that the strategy is a firm basis for further research in the area of the self-learning adaptive coordinated traffic corridor control.

Д-р **КОСТАНДИНА ВЕЉАНОВСКА**
Е-пошта: kostandina@rocketmail.com
Универзитет „Св. Климент Охридски“
Факултет за администрација и менаџмент на информациски системи
Партизанска бб, 7000 Битола, Република Македонија
Д-р **КРИСТИ БОМБОЛ**
Е-пошта: kristi.bombol@uklo.edu.mk
Универзитет „Св. Климент Охридски“, Технички факултет
П. Фах 99, 7000 Битола, Република Македонија
Dr. **TOMAŽ MAHER**
E-mail: tmaher@fgg.uni-lj.si
Univerza v Ljubljani, Fakulteta za gradbeništvo in geodezijo
Prometnotehniški inštitut
Jamova 2, 1000 Ljubljana, Republika Slovenija

АПСТРАКТ

ТЕХНИКАТА НА ПРИНУДНО УЧЕЊЕ ВО ПРОЕКТИРАЊЕТО НА СООБРАЌАЈНАТА КОНТРОЛА НА АВТОПАТ СО ПОВЕЌЕ ПРИСТАПИ

Соодветно проектирана контрола на пристап на автопат може да го намали вкупното време на патување во системот за 30% и последователно да ја зголеми безбедноста на влевање на возилата. Досегашните зависни системи за контрола на сообраќајот на пристапот на автопат беа од локален или регулаторски тип, што значи дека не беа целосно адаптивни во вистинското значење на зборот.Оттука, на сообраќајниот ток може да се влијае со бројни техники на интелигентните транспортни системи (ИТС).

Во овој труд е претставен современ пристап кон филозофијата на оптимална и адаптивна контролна стратегија на повеќепристапен автопат со затворена јамка. Предложената методологија ја користи техниката на вештачка интелигенција, позната како принудно учење (ПУ) со повеќекратни агенти и го применува алгоритмот на Q-учење.

Проектирана беше една делница од мрежата на автопат со три ленти во секоја насока и три пристапа (влезни рампи). Беа поставени детектори и светлосни сообраќајни знаци на сите три рампи. Како влезни променливи во моделот беа земени: големината на сообраќајни токови, густината на сообраќајот на главниот коридор и сообраќајната побарувачка на пристапите кон автопат. Излезните променливи величини се однесуваа на брзината на патување на коридорот, на вкупното време на патување и на вкупното време на застанување. За да се реализира техниката на ПУ, се примени микросимулаторот VISSIM и директното програмирање на симулаторските функции. За време на симулација е избран врвниот час.

Моделот беше тестиран во две фази. Неговата ефикаснот беше споредена со ALINEA. Се заклучи дека предложената стратегија може да одговори на динамичките влезови од сензорите на околината и на динамичкипроемнливата околина. Контролата политика самата се менуваше како одговор на промените на инхерентните системски карактеристики. Се потврди дека стратегијата е вистински адаптивна и зависна од сообраќајната побарувачка на коридорот во реално време.

КЛУЧНИ ЗБОРОВИ

Пристап на автопат, сообраќајни токови, контрола, стратегија, вештачка интелигенција, Q- учење, симулација

LITERATURE

[1] **Sutton, R.S.**, and **Barto, A.G.**, *Reinforcement Learning - An Introduction*. MIT Press, Cambridge, Massachusetts, 1998

[2] **Karakoulas, G.J.**, *Probabilistic Exploration in Planning while Learning*. In Proceedings of the *Eleventh International Conference on Uncertainty in Artificial Intelligence*, 1995

[3] **Abdulhai, B.**, **Pringle, R.**, and **Karakoulas, G. J.**, *Reinforcement Learning for ITS: Introduction and a Case Study on Adaptive traffic Signal Control*, TRB, Washington D.C., 2001

[4] **Veljanovska, K.**, *Development of an Optimal Adaptive Control Strategy for Motorway Access Control* (PhD Dissertation, Faculty of Technical Sciences, Sv. K. University of Ohrid, Bitola, 2008 (In Macedonian)

[5] **Abdulhai, B.**, *Fundamentals of ITS and Traffic Management*, Lecture notes, University of Toronto, Toronto, Canada, 2002

[6] **Abdulhai, B.**, **Pringle, R.** and **Karakoulas, G. J.**, *Reinforcement Learning for ITS: Introduction and a Case Study on Adaptive Traffic Signal Control*, Transportation Research Board, Washington D.C., 2001

[7] **Abdulhai, B.**, **Pringle, R.** and **Karakoulas, G.J.**, *Reinforcement Learning for True Adaptive Traffic Signal Control*, ASCE Journal of Transportation Engineering, Volume 129, Number 3, pp278-285, 2003

[8] **Adler, J.L.** and **Blue, V.B.**, *A cooperative multi-agent transportation management and route guidance system*, Transportation Research Part C: Emerging Technologies Vol: 10 Issue: 5-6, 2002

[9] **Bogenberger, K.** and **May, A.**, *Advanced Coordinated Traffic Responsive Ramp Metering Strategies*, Institute of Transportation Studies, California Partners for Advanced Transit and Highways (PATH) University of California, Berkeley, p.n. 30, 1999

[10] **Chang, G.L.**, **Wu, J.** and **Cohen, S.**, *Integrated real-time metering model for non-recurrent congestion: framework and preliminary results*. Transportation Research Record: Journal of the Transportation Research Board, No. 1446, TRB, National Research Council, Washington, D. C., 1994

[11] **Hadj-Salem, H.**, **Blosseville, J.M.** and **Papageorgiou, M.**, *ALINEA: A Local Feedback Control Law for On-ramp Metering*; A Real-Life Study, Third International Conference on Road Traffic Control, IEEE, Stevenage Printing Limited, Great Britain, pp. 194-198, 1990

[12] **Hasan, M.**, **Uha, M.**, **Ben-Akiva, M.**, *Evaluation of Ramp Control Algorithms Using Microscopic Traffic Simulation*, Transportation Research Part C, Vol. 10, No.3, 2002, pp. 229-256

[13] **Huhns, M.** and **Stephens, L.M.**, *Multiagent systems: a modern approach to distributed artificial intelligence*, Multiagent Systems and Societies of Agents, MIT Press, Cambridge, MA, USA, pp. 79-120, 1999

[14] **Kotsialos, A.**, **Papageorgiou, M.**, **Middelham, F.**, *Optimal Coordinated Ramp Metering with Advanced Motorway Optimal Control*, Transportation Research Record 1748, pp. 55-65

[15] **Kwon, E.**, **Nanduri, S.**, **Lau, R.** and **Aswegan, J.**, *Comparative Analysis of Operational Algorithms for Coordinated Ramp Metering*, Transportation Research Record 1748, pp 144-152

[16] **Mitchell, T.**, *Machine learning*, McGraw-Hill, New York, USA, 1997

[17] **Papageorgiou, M.**, *ALINEA: A Local Feedback Control Law for On-Ramp Metering*, Transportation Research Record 1320, 1990

[18] **Papageorgiou, M.**, et al., *ALINEA local ramp metering - Summary of field results*. Transportation Research Record: Journal of the Transportation Reserch Board, No. 1603, TRB, National Research Council, Washington, D.C., 1998, p. 90-98

[19] **Smaragdis, E.** and **Papageorgiou, M.**, *A Series of New Local Ramp Metering Strategies*. Transportation Research Board, 82 nd Annual Meeting, Washington, DC, 2003

[20] **Stephanedes, Y.**, **Chang, K.**, *Optimal Ramp-metering Control for Freeway Corridors*, Applications of Advanced Technologies in Transportation Engineering - Proceedings of the Second International Conference, 1991, pp. 172-176

[21] **Wei, C.**, *Analysis of Artificial Neural Network Models for Freeway Ramp Metering Control*, Artificial Intelligence in Engineering, Vol 15, Elsevier, 2001

[22] **Zhang, H.M.** and **Ritchie, S.G.**, *Freeway Ramp Metering Using Artificial Neural Networks*, Transportation Research Part C, Vol. 5, No. 5, Elsevier Science Ltd, Great Britain, 1997, pp. 273-286