Tunnel Ventilation Control Using Reinforcement Learning
Methodology∗

Baeksuk CHU∗∗, Dongnam KIM∗∗, Daehie HONG∗∗∗, Jooyoung PARK∗∗∗∗, Jin Taek CHUNG∗∗∗ and Tae-Hyung KIM†

The main purpose of tunnel ventilation system is to maintain CO pollutant concentration and VI (visibility index) under an adequate level to provide drivers with comfortable and safe driving environment. Moreover, it is necessary to minimize power consumption used to operate ventilation system. To achieve the objectives, the control algorithm used in this research is reinforcement learning (RL) method. RL is a goal-directed learning of a mapping from situations to actions without relying on exemplary supervision or complete models of the environment. The goal of RL is to maximize a reward which is an evaluative feedback from the environment. In the process of constructing the reward of the tunnel ventilation system, two objectives listed above are included, that is, maintaining an adequate level of pollutants and minimizing power consumption. RL algorithm based on actor-critic architecture and gradient-following algorithm is adopted to the tunnel ventilation system. The simulations results performed with real data collected from existing tunnel ventilation system and real experimental verification are provided in this paper. It is confirmed that with the suggested controller, the pollutant level inside the tunnel was well maintained under allowable limit and the performance of energy consumption was improved compared to conventional control scheme.

Key Words: Tunnel Ventilation Control, Reinforcement Learning (RL), Actor-Critic Architecture, Gradient-Following Algorithm

1. Introduction

An appropriate operating of roadway tunnel ventilation system provides the drivers passing through the tunnel with comfortable environment and safe driving condition. At the same time, the tunnel ventilation system consumes large amount of energy. So, it is desired to have an efficient operating algorithm for the tunnel ventilation in the aspects of safe and comfortable driving environment as well as saving energy. The main target of the roadway tunnel ventilation is to maintain CO pollutant and VI (visibility index) to a certain level. CO is mainly emitted from gasoline passenger cars. The amount of CO pollutant that is over an allowable level may cause a fatal injury to human body. Generally, 100 ppm is the maximum CO limit that can be allowed. VI is mainly decreased by the smoke emitted from diesel buses and trucks. The low VI may considerably decrease drivers’ driving capability due to poor visibility and even induce traffic accidents.

The pollutants in the tunnel are exhausted from passing vehicles, which are moving sources. Moreover, their transient behavior is characterized with time delay. Due to such problems, complex and nonlinear system like tunnel ventilation is difficult to control with conventional quantitative methods. The most popular control method for such systems is fuzzy logic control. There have been many studies for tunnel ventilation control using fuzzy logic. However, the tunnel ventilation control using fuzzy control method has several problems. First, it is dif-
difficult to build rule database which is necessary for fuzzy inference process in tunnel ventilation system, because it is hard to obtain and accumulate tunnel operators’ or experts’ knowledge. Second, it is not easy to determine an appropriate membership function for fuzzy controller. So, the previous studies listed above designed the membership function by relying on simple experience and using trial and error method.

To overcome such problems, the control algorithm used in this research is reinforcement learning (RL) method. RL is a goal-directed learning of a mapping from situations to actions without relying on exemplary supervision or complete models of the environment. The goal of RL is to maximize a reward or reinforcement signal which is an evaluative feedback from the environment. In the process of constructing reward of the tunnel ventilation system, maintaining pollutant concentration level under an allowable limit is the most important purpose. Besides it, energy consumption is also a considerable factor to be included in the reward formulation. Consequently, the controller used in this study is designed to optimally satisfy both control objectives through the learning process of RL.

RL has been an active research area in machine learning, control engineering, and so on\(^{(6)-(9)}\). Among many categories of RL, this study is based on a combination of two algorithms. The one is actor-critic algorithm\(^{(6)}\) which is sometimes called adaptive-heuristic-critic (AHC) learning architecture. The other is gradient-following algorithm. In actor-critic learning structure, the controller is divided into two components, the critic (evaluation) module and the actor (control) module. Both modules have their own learning process based on gradient-following algorithm, respectively. While most of previous researches about RL are concentrated on narrow application area confined to a typical example such as inverted pendulum, this research shows that RL can be also implemented to various real-world systems.

This paper is organized as followings. In section 2, the target tunnel ventilation system of this research is briefly introduced. In section 3, the basic concepts of RL and actor-critic algorithm based on gradient-following algorithm are demonstrated for tunnel ventilation controller design. Then, in section 4, the simulation results performed with real data collected from the target tunnel system and the experimental results are shown. It is confirmed that the RL-based controller shows efficient performance in the aspects of both maintaining pollutant concentration level under an allowable limit and saving energy consumption. Finally, the last section contains concluding remarks.

2. Tunnel Ventilation System

The target tunnel for this research is Dunnae tunnel located in Youngdong highway, in Korea. Figure 1 and Table 1 show the schematic diagram and detail specifications of the tunnel. To observe the pollutant levels, CO sensor and VI sensor are installed inside the tunnel with an appropriate interval. Traffic counter located at tunnel entrance records the number of cars entering into the tunnel. In order to ventilate the pollutants, total 32 jet-fans are installed on the ceiling.

The distribution of pollutants inside the tunnel is usually represented with one-dimensional diffusion-advection equation\(^{(3)}\). It has pollutant inputs from passing vehicles as a source term.

\[
\frac{\partial c}{\partial t} = \frac{\partial}{\partial x} \left( k \frac{\partial c}{\partial x} \right) - V_w \frac{\partial c}{\partial x} + q
\]  

(1)

In Eq. (1), \( c \) indicates the pollutant concentration existing inside the tunnel. \( V_w \) and \( k \) are wind velocity and diffusion coefficient, respectively. The first term on the right-hand side explains the diffusion of pollutants and the second term does the advection by the wind. The pollutant source \( q \) increases the pollutant level inside the tunnel. However, because the advection and source term generally domi-

<table>
<thead>
<tr>
<th>Table 1 Specifications of Dunnae tunnel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tunnel</strong></td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td>Width</td>
</tr>
<tr>
<td>Height</td>
</tr>
<tr>
<td>Lanes</td>
</tr>
<tr>
<td>Cross-sectional area</td>
</tr>
<tr>
<td>Ventilation</td>
</tr>
</tbody>
</table>

Fig. 1 Schematic diagram of Dunnae tunnel with jet-fans
nate the pollutant distribution, the diffusion term is often ignored. Then, the one-dimensional advection equation in which the diffusion term is canceled is like following Eq. (2).

\[
\frac{\partial c}{\partial t} = -V_u \frac{\partial c}{\partial x} + q
\]  

(2)

In order to estimate the change of pollutant distribution, it is necessary to identify the wind velocity inside the tunnel. It can be calculated by force balance equation, that is, using Newton’s second law,

\[
\rho A \frac{dV_u}{dt} = \sum F
\]

\[
\sum F = F_t + F_j + F_r + F_n
\]  

(3)

where \( \sum F \) is the summation of the forces that affect the wind in its flow velocity inside the tunnel. \( F_t \) is the traffic ventilation force by passing vehicles, \( F_j \) is the equipment ventilation force by jet-fan operation, \( F_r \) denotes the wall friction resistance and fluent loss at the entrance and exit, and \( F_n \) explains the wind resistance by the natural wind outside the tunnel. Besides, \( \rho \) is the density of air, \( A \) is the cross-sectional area of the tunnel, and \( L \) is the longitudinal length of the tunnel.

3. Reinforcement Learning Using Actor-Critic Architecture

In this section, the RL algorithm using actor-critic architecture and gradient-following algorithm are demonstrated to solve the learning control problems. It draws control efforts from the distribution characterized by some parameters. The adjustment of the parameters is based on gradient-following algorithm with an appropriate step size.

3.1 Actor-critic algorithm

The actor-critic model, which is sometimes called adaptive-heuristic-critic (AHC) learning architecture, includes two principal components, the critic (evaluation) module and the actor (control) module. The actor is used to generate optimal control actions according to a certain policy. The critic is used to evaluate the policy represented by the actor and to provide the action and the critic itself with the evaluation information. The critic and actor module gradually converge to the optimal performance with its own learning processes based on stochastic approximation, respectively. In this research, they are determined by a class of gradient-following algorithm.

Like other RL methodologies, actor-critic algorithm is distinguished from other computational approaches by emphasizing on learning from direct interaction with its environment, without relying on exemplary supervision or complete models of the environment. Figure 2 gives an overview of actor-critic algorithm. The adaptive critic receives the state vector and the external reinforcement signal, the reward which is an evaluative feedback, from the environment as inputs, and transforms them into the internal heuristic reinforcement signal, TD signal. Using the information from the critic and environment, the actor outputs the control actions that tend to increase the long-run sum of the reward.

The goal of RL method is to obtain the actor that maximizes the expected reward accumulation toward the future, which can be represented via the concept of value function. For a given state \( s \in S \) and action \( u \in U \) at each time \( t \in [0, 1, 2, \ldots] \), the value function is defined as

\[
V^\pi(s) = E \left\{ \sum_{i=0}^{\infty} \gamma^i r_{t+i} \bigg| s_0 = s, \pi \right\} \]  

(4)

where \( \pi \) denotes the current policy \( \pi(u|s) = Pr[u_t = u|s_t = s] \) from which actions are determined in each state, \( \gamma \) is the discount factor with \( \gamma \in [0, 1] \), and \( r_t \) is the reward from the interaction with its environment. Also, the evaluation of the prospect of a command \( u \) for a given state \( s \) is usually performed by computing another important concept, the action-value function:

\[
Q^\pi(s, a) = E \left\{ \sum_{i=0}^{\infty} \gamma^i r_{t+i} \bigg| s_0 = s, u_t = a, \pi \right\} \]  

(5)

in which the value according to each action is expressed as a function of state and action. The former value function can be formulated by averaging over all actions in each state of the action-value function.

3.2 Actor module

The action selection policy of actor module is estimated using a function approximator with its own parameters. For each state \( s \), it is designed to be a stochastic unit \( \pi_\theta(u|s) \) that computes its output \( u \) as a stochastic density function parameterized by the vector \( \theta \). Let’s define the objective function \( J(\theta) \) to be maximized as the value function from a designated start state \( s_0 \):

\[
J(\theta) = V^\pi(s_0) = \int_U \pi_\theta(u|s_0)Q^\pi(s_0, u)du \]

\[
= \int_S d^\pi(s) \int_U \pi_\theta(u|s)r(s, u)du ds \]  

(6)

where \( r(s, u) \) denotes the reward, and \( d^\pi(s) = \sum_{i=0}^{\infty} \Pr[s_t = s_0] \).
$s|s_0, π_θ|$ is the discounted state distribution starting at $s_0$ and then following $π_θ$.

In order to update the policy parameters of the actor module, ‘policy gradient’ approach is applied, in which the policy parameters are adjusted using the steepest gradient ascent for the optimal policy with respect to the objective function $J(θ)^{(12)}$:

$$θ_{t+1} = θ_t + βθ∇_θ J(θ_t)$$

(7)

where $βθ$ is an appropriate step size of actor. The main issue of this scheme is to estimate the gradient information $∇_θ J(θ_t)$ with respect to the policy parameter vector $θ$. Making use of the ‘policy gradient theorem’^(12), (13)^, the true policy gradient can be estimated like follows:

$$∇_θ J(θ) = ∇_θ \left( ∫ s \int d^θ(s) ∫ u π_θ(u|s) r(s,u) du ds \right)$$

$$= ∫ s d^θ(s) ∫ u π_θ(u|s)(Q^θ(s,u) - b^θ(s)) du ds$$

(8)

where $b^θ(s)$ denotes an arbitrary function called a baseline. Because $∫ u π_θ(u|s) du = 0, \forall s \in S$, it does not affect the gradient estimate as some bias. However, taking $b^θ(s)$ equal to an approximation of value function $V(s)$ induces the effect to minimize the estimate’s variance^(14). Using the value function approximator to reduce the variance of gradient estimate is critical for rapid learning of actor-critic algorithm.

So, during the state transition from $s_t$ to $s_{t+1}$ by action $u_t$, the update formulation of the policy parameters can be written as

$$θ_{t+1} = θ_t + βθ∇_θ J(θ_t)$$

$$= θ_t + βθ∇_θ log π_θ(u_t|s_t)(Q^θ(s_t,u_t) - V(s_t))$$

(9)

through a stochastic sampling. If the above can be achieved, then $θ$ can usually be assured to converge to a locally optimal policy in the performance measure $J(θ)^{(12)}$. With the immediate reward $r_t$, some estimator of $Q^θ(s_t,u_t)$ can be chosen as $r_t + γV(s_{t+1})^{(12), (15)}$, then Eq. (9) is reformulated like following:

$$θ_{t+1} = θ_t + βθ∇_θ log π_θ(u_t|s_t)$$

(10)

$$= θ_t + βθ∇_θ log π_θ(u_t|s_t)$$

where $δ_t$ is the temporal difference (TD):

$$δ_t = r_t + γV(s_{t+1}) - V(s_t)$$

(11)

In this paper, normal distribution is employed as the density $π_θ(u|s)$ that governs the control selection. The actual action $u$ is chosen by exploring a range around the mean point which is determined by a parametric approximator $k_θ(s)$ weighted by vector $θ$. This range of exploration corresponds to the variance of the probability distribution. So, the actual output of the stochastic unit can be set as

$$u ∼ N(k_θ(s), σ^2)$$

(12)

And $u$ is a Gaussian random variable with a density function:

$$π_θ(u|s) = \frac{1}{σ\sqrt{2π}} \exp \left( \frac{-(u-k_θ(s))^2}{2\sigma^2} \right)$$

(13)

where the variance $σ$ is constant. Using normal distribution as action selection policy, the update of the policy parameters, Eq. (10) is reconfigured like following:

$$θ_{t+1} = θ_t + βθ δ_t \frac{d log π_θ(u_t|s_t)}{d k_θ(s)} \frac{dk_θ(s)}{dθ}$$

$$≈ θ_t + βθ δ_t \frac{u_t - k_θ(s_t)}{σ^2} \frac{dk_θ(s)}{dθ}$$

(14)

### 3.3 Critic module

The critic generates an estimate of the value function and provides the actor and critic itself with the information through the form of TD signal. It helps the actor and critic itself to update their parameters toward an optimal performance. So, an accurate and fast estimation of the value function is very significant. RL in continuous state and action spaces like many realistic applications needs a function approximator such as neural networks to approximate the value function. Employing a linear function approximator with following basis vector:

$$φ(s) = [φ_1(s), φ_2(s), …, φ_N(s)]^T,$$

(15)

the estimated value function approximator can be represented as

$$V(s; w) = φ^T(s) · w$$

(16)

which is parameterized by the weight vector $w$.

To approach the accurate estimation, the value function approximator should be updated according to gradient-following scheme such as

$$Δw ≈ \frac{1}{2} \frac{d}{dw} [V^θ(s_t) - V(s_t; w)]^2$$

(17)

If we select $v_t ≡ r_t + γ V(s_{t+1}; w_t)$ as an estimator of $V^θ(s_t)$ during the state transition from $s_t$ to $s_{t+1}$ by action $u_t$, Eq. (17) can be converted to

$$\frac{1}{2} \frac{d}{dw} [v_t - V(s_t; w)]^2$$

$$= \frac{1}{2} \frac{d}{dw} [r_t + γ V(s_{t+1}; w_t) - V(s_t; w_t)]^2$$

$$= \frac{d}{dw} \left[ γ \frac{dV(s_{t+1}; w_t)}{dw} - \frac{dV(s_t; w_t)}{dw} \right]$$

(18)

Combining Eq. (18) to Eq. (17), the update formulation of the value function parameters can be obtained as

$$w_{t+1} ≈ w_t + βθ δ_t \frac{dV(s_t; w_t)}{dw}$$

(19)

where $βθ$ is a step size of critic module. This update procedure is the famous update for the temporal difference learning algorithm, TD($λ$), with $λ = 0$. Note that
by neglecting the contribution of the $V(s_{t+1}; w_t)$ term to the gradient as in Eq. (18), the convergence rate of the value function update can be considerably increased in practice, which is supported by a significant experimental evidence(8).

4. Simulations and Experimental Results

Most of past research about RL has been confined to on narrow applications, such as inverted pendulum balancing or acrobot swing-up control. They were model problems for performance comparison of some theoretical improvements. However, as various RL algorithms that show sufficiently good performance have been developed, it is time to make a progress not only in theoretical studies but also in real-world applications. This paper explores the feasibility for applying the RL to real systems. The control algorithm proposed in this paper is verified with computer simulations and real experiments. The data for the simulations were gathered from the real tunnel system, Dunnae Tunnel located in Youngdong highway in Korea. The states measured by sensors consist of CO pollutant level, VI, and pollutant emission rate by passing vehicles. In order to simplify the descriptions in this paper, only the CO level and pollutant emission rate will be considered in the controller design. Adding VI level to the control algorithm is quite straightforward.

To solve the continuous state space problem in RL, a linear function approximator is used for value function estimator in the critic module and for output unit in the actor module. It is designed as a linear combination of three components parameterized by a weight vector. The first component of the approximator is the difference of the CO sensor feedback from an allowable reference CO pollutant level, 25 ppm in this research. The second component is the difference between average reference emission rate and currently observed emission rate. The third component is a bias term. In addition, the control output of the proposed algorithm is the relative number of running jet-fans to the nominal number of which the jet-fans are operated under the condition of nominal pollutant level. The total number of jet-fans which can be driven is 32 and the nominal number is chosen as 15.

The reward formulation is a main criterion to RL process and an important connection between the control algorithm and the system. The reward reflects the objective to be achieved by controller and a penalty for violating a constraint of the system. In this study, the reward has been constructed with combination of pollutant reduction term as the objective and energy consumption term as the constraint. In Eq. (20), the pollutant level over an allowable limit and the energy consumption proportional to the number of running jet-fans are combined with an appropriate weighting factor, $K$.

$$
\text{reward} = \begin{cases} 
-((CO_{current} - CO_{ref}) + K \cdot E_{JF}), & \text{if } CO_{current} > CO_{ref} \\
-E_{JF}, & \text{if } CO_{current} < CO_{ref} 
\end{cases}
$$

where $CO_{ref}$ is the allowable reference CO pollutant level, 25 ppm, $CO_{current}$ is the current CO sensor feedback, and $E_{JF}$ is the energy consumed by the operation of jet-fans.

The parameters for the control algorithm are set as followings. The discount factor $\gamma$ is 0.9. The learning rates for the actor network, $\beta^\theta$, and critic network, $\beta^V$ are set to 0.1, which is manually optimized. The variance of the actor probability distribution is 1.5. The initial weights of actor and critic module are all zeros. A time step is 1 min.

4.1 Simulations

Simulations are run for 5 000 time steps which are 10 repeated replicas of 500 real sample data. Figures 3 and 4 show the peak value of CO inside the tunnel for the case that no ventilation control is applied. The 2-D plot in Fig. 3 contains the whole 5 000 time steps and the 3-D plot in Fig. 4 shows detail distribution of CO pollutant along the longitudinal distance of the tunnel for the last 40 time steps. In this case, the pollutant emission by passing vehicles is the only input source to the system. In other words, any control input except the operation of the nominal number of jet-fans is not conducted to ventilate CO...
the tunnel. As such, it is shown that the maximum CO pollutant level considerably exceeds 25 ppm, the control objective.

If a control input based on the RL algorithm proposed by this study is added to the system, the pollutant concentration decreases like Figs. 5 and 6. Figure 5 also shows a learning process of the RL-based controller. As the learning progresses, the controller raises up the number of running jet-fans, such that the CO level is maintained under the allowable limit, 25 ppm. On the other hand, if the CO pollutant is maintained well below the allowable limit and excessive energy is consumed by running unnecessary overworking jet-fans, the controller decreases the number of jet-fans and saves the energy consumption. These two facts from the figures explain that the RL-based controller appropriately follows the control objectives expressed by the reward formulation for this system. After a sufficient time is spent for learning, as shown in Fig. 6, the CO pollutant level along time axis stays near the allowable limit and the energy consumption becomes very efficient. However, it is shown in the figures that the peak of the pollutant exceeds the allowable limit at several locations. This situation happens when the number of vehicles entering the tunnel suddenly increases a lot more than normal level. Such a sudden rush of vehicles inevitably induces the excess over the allowable pollutant level due to the two control objectives, both controlling pollutant level under a allowable limit and minimizing energy consumption. If the control objective is just to reduce the peak value of pollutant absolutely under 25 ppm, it can be achieved by setting the reference value to a lower reference CO level than 25 ppm. In Fig. 7, the reference CO level is set to 23 ppm, so all the maximum peak values of CO are under 25 ppm. In this case, the increase of energy consumption is naturally accompanied.

The performance of the proposed control method is evaluated with respect to the reduction of CO pollutant level and energy consumption. Table 2 shows the mean value, standard deviation, maximum/minimum value of peak CO level and consumed energy during the last 500 samples. Among three cases, the uncontrolled case corresponds to Figs. 3 and 4 and the controlled case by RL when the objective CO level is 25 ppm does to Figs. 5 and 6. These two cases have the similar mean value of CO level but the maximum CO level of the controlled case by RL is lower than the uncontrolled case. In addition, the energy consumption of the controlled case is also lower. These results show that the RL-based controller nicely achieves the two control objectives of this system, the pollutant reduction and low energy consumption by an efficient opera-

<table>
<thead>
<tr>
<th>Control</th>
<th>CO_{mean} (ppm)</th>
<th>CO_{std} (ppm)</th>
<th>CO_{max} (ppm)</th>
<th>CO_{min} (ppm)</th>
<th>Energy (kWh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncontrolled case (constant operation of nominal number of jet-fans)</td>
<td>24.48</td>
<td>2.2355</td>
<td>29.71</td>
<td>19.78</td>
<td>3750</td>
</tr>
<tr>
<td>Controlled case by RL (when objective CO level is 25 ppm)</td>
<td>24.74</td>
<td>0.6874</td>
<td>26.86</td>
<td>22.45</td>
<td>3491.7</td>
</tr>
<tr>
<td>Controlled case by RL (when objective CO level is 23 ppm)</td>
<td>22.48</td>
<td>0.9289</td>
<td>24.97</td>
<td>19.90</td>
<td>4562</td>
</tr>
</tbody>
</table>
4.2 Experimental results

The suggested algorithm in the research is once more verified through real experiment. The controller is based on PC and an additional communication module is devised for the interface between the controller and the field system. Figure 8 shows the outline of the tunnel ventilation experiment with the experimental apparatus. Most of conditions of the experiment are identical to the simulations, but some of them are properly adjusted according to the filed circumstances. In this experiment, the objective CO pollutant level is chosen to be 12 ppm. As seen in the simulations, the time demanded for learning is not too short. So, in order to reduce the learning time in real experiment, the learning parameter which is found from computer simulations is used as initial value. In this way, we can also easily determine the initial learning parameter for other tunnels.

Figure 9 shows the first 30 time steps when the learning process has just begun. The figure located in the top among three subplots is about operation step in which the number of jet-fans is graded into total 10 steps. The second plot in the middle shows the internal wind velocity, and the last plot in the bottom does the CO pollutant level. In this case, the current CO level considerably exceeds the objective, 12 ppm. Because it is the early step of learning process, the desired control performance is not observed yet. On the contrary, Fig. 10 depicts the 30 time steps after the learning has proceeded for some time. In this figure, the CO pollutant level is maintained near to the objective, 12 ppm by an appropriate operation of the controller. In addition, since there is not an excessive overworking of control elements, the energy consumption is efficiently minimized. Therefore, the two control objectives of the system are sufficiently achieved.

5. Concluding Remarks

In order to control tunnel ventilation system efficiently, reinforcement learning method based on actor-critic architecture and gradient-following algorithm is used in this paper. The two objectives of tunnel ventilation are composed of maintaining pollutant concentration level under an appropriate limit and minimizing energy consumption used to operate control elements. By importing the two goals into the reward formulation of RL, tunnel ventilation controller was designed to produce the optimal control input with respect to all of the two aspects. The proposed controller was verified through various simulations and real experiments and their analyses were performed. It is confirmed that the RL-based controller shows efficient performance in the aspects of both managing appropriate pollutant concentration level and saving energy consumption. While most of previous research about RL was confined to some scholarly examples, this research opens the possibility of applying it to a real system.

Acknowledgment

This work was performed in part of 2003 industrial-educational cooperation project, ‘Underground, Fire, and Environment Research’ supported by Korea Institute of Construction and Transportation Technology Evaluation and Planning.
References


