2020 SICE International Symposium on Control Systems (SICE ISCS)
Tokushima University, Tokushima, Japan, March 3-5, 2020

Airflow Direction Control of Air Conditioners Using Deep Reinforcement Learning

Yuiko Sakuma[1†] and Hiroaki Nishi[2]

[1]Graduate School of Science and Technology, Keio University, Yokohama, Japan
(Tel : +81-45-566-1785; E-mail: sakuma@west.sd.keio.ac.jp)
[2]Department of System Design, Faculty of Science and Technology, Keio University, Yokohama, Japan
(Tel : +81-45-566-1785; E-mail: west@sd.keio.ac.jp)

Abstract: Achieving a uniform comfort within an indoor housing environment is important for health and productivity while saving the energy consumption of a Heating, Ventilation, and Air Conditioners (HVAC) device. The optimal control of an HVAC system is a well-studied area. While many works explore the optimal temperature set-point, a few works consider effective airflow direction control. This work proposes an airflow direction control method that aims uniform comfort of the indoor environment using a deep reinforcement learning (DRL) approach. We implemented our proposed DRL framework using computational fluid dynamics (CFD) simulation software. Our proposed method was evaluated for comfort and energy consumption. The experimental results show the improvements for our proposed method in comfort by 21.3 % while reducing energy consumption by 34.5 % for the average than the baseline method.

Keywords: HVAC control, Deep reinforcement learning

1. INTRODUCTION

Comfort is important for health and productivity to meet the increased demand for a high-quality indoor environment. Especially, uneven heating within a room induces thermal discomfort. Placement of obstacles and changing solar radiation patterns can be the examples which induce uneven temperature distribution. At the same time, along with a growing awareness of environmental problems, reducing the energy consumption of the residential sector has been an important issue. In the US, buildings are responsible for 40% of total consumption and Heating, Ventilation, and Air Conditioners (HVAC) systems account for 48 % of the total energy consumption of homes [1]. Therefore, an HVAC control method that balances comfort and energy consumption is required.

While many works assume the uniform comfort within a room, several works consider the zone-level control. Agarwal et al. [2] proposed an ON-OFF strategy of the HVAC system with multiple diffusers. They controled the HVAC system according to the occupancy information. Ghahramani et al. [3] controlled zone temperature set-points using personalized thermal comfort preferences. Abedi et al. [4] further explored the effectiveness of flexible control of the air conditioner (AC). They adjusted the airflow direction by considering occupants location which resulted to average 59% energy savings and achieved desired thermal comfort. However, control of the airflow direction of an AC to achieve the uniform comfort within a room has not been proposed.

This paper proposes an airflow direction control method of ACs that achieves uniform comfort within a room. While many works that study HVAC control methods exist, there are two approaches. One category uses model predictive control (MPC). MPC uses a system model to predict future indoor environments to decide control actions [5]. The control performance is dependent on the modeling accuracy. Modeling often requires detailed information of the room such as materials and floor plans. Therefore, modeling complexity is high under the existence of obstacles or complex room shapes. Authors of [6] and [7] proposed simplified models to reduce modeling complexity. In our case of controlling the airflow direction, modeling of air mixing is complex and may reduce control performance.

To cope with the limitation of MPC that require accurate thermal models, data-driven methods are proposed. Recent works propose a reinforcement learning (RL)-based controller that uses real-time data inputs for HVAC control. Barrett et al. [8] used Q-learning [9] to optimize both occupant comfort and energy costs. Q-learning often suffers from slow convergence because of low sample efficiency. Li et al. [10] proposed a coarse model that achieves a fast converge of Q-learning for the HVAC controller. However, Q-learning also has a limitation for managing a large size system state.

Recently proposed deep reinforcement learning (DRL) techniques show good performance for managing a large size system state and complex control problems as shown successfully in playing Atari [11] and AlphaGo games [12]. Several works show the effectiveness of DRL for HVAC control. We propose the DRL-based airflow direction control method. The main contributions of this paper are as follows:

- We propose a DRL-based controller that achieves the uniform comfort within a room by considering airflow direction. We designed the system states, reward function, and control action for the controller that enables learning of the optimized actions.
- We implemented the simulation for training and evaluation using the computational fluid dynamics (CFD) software and real-world weather information.
- Our proposed method was evaluated against a fixed and rule-based controller for rooms without and with partitions.

† Yuiko Sakuma is the presenter of this paper.

978-4-907764-64-7

This paper is organized as follows. Section 2 summarizes the related works. Our proposed DRL based controller is described in Section 3. The experimental design is presented, and the experimental results are discussed in Section 4. Section 5 summarizes and concludes the paper.

2. RELATED WORKS

2.1 Deep Q-learning

While various RL methods are proposed, they can be classified into two categories; model-based and model-free algorithms. Model-based algorithms are used when the parameters of the Markov decision process (MDP) are known and the model of the environment can be developed. Since estimating the indoor environment model is difficult in our case, we focus on the model-free approach.

The model-free approach directly optimizes the reward function from the state. Q-learning is one of the most commonly used algorithms. Q-learning algorithm learns the value function $Q(s_t, a_t)$ for each state s_t and action a_t in an online manner. Deep Q-Network (DQN) [11] uses the artificial neural network to approximate the Q-value that enables to manage a large state-action space. It uses separating Q-networks of main Q-network (Q_m) and target Q-network (Q_t) to stabilize learning. After calculating $Q_m(s_t, a_t)$, the action is chosen by using the ε-greedy policy. Taking the selected action a_{t+1}, the next state s_{t+1} is observed and the immediate reward r_{t+1} is calculated. Then Q_m is updated as in Eq. (1) and to approximate$Q_m(s_t, a_t)$, the loss $E(s_t, a_t)$ is calculated using mean squared error ($E(s_t, a_t)$) as shown in Eq.(2).

$$Q_m(s_t, a_t) = Q_m(s_t, a_t) + \eta\left(r_{t+1} + \gamma max_a Q_t(s_{t+1}, a) - Q_m(s_t, a_t)\right) \quad (1)$$

$$E(s_t, a_t) = \left(r_{t+1} + \gamma max_a Q_t(s_{t+1}, a) - Q_m(s_t, a_t)\right)^2 \quad (2)$$

2.2 DRL-based HVAC control

While several works apply classic RL algorithms such as Q-learning ([8], [10], and [13]) and SARSA ([14]) for HVAC control tasks, many recent works use DRL algorithms. Nagy et al. [15] compared rule-based control (RBC), MPC, model-based RL, and model-free RL techniques. Although the result showed that MPC and model-based RL outperform model-free RL, their proposed model-free RL controller showed the benefits of faster computation time and robustness on unexpected changes in the indoor environment. They showed that model-free RL techniques can be a solution when perfect knowledge about the indoor environment is not available. Yang et al. [16] used DQN for the HVAC controller to control the temperature set-point and showed that it outperforms RBC by over 10%. Wei et al. [17] also used a DQN-based controller to control the temperature set-point. Further, they heuristically adapted their proposed algorithm for multi-zone HVAC control. They showed that the DQN-based controller can also be used when multiple zones thermally affect each other. Therefore, we apply the DQN algorithm to our proposed controller.

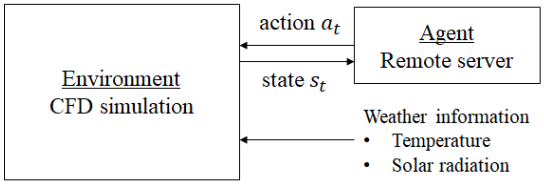


Fig.1 DRL control framework.

While DQN is a value-based algorithm, Gao et al. [18] used a policy-based algorithm, Deep Deterministic Policy Gradients (DDPG) [19] to deal with a continuous action space. However, this is out of the scope of our work because we consider a discrete action space. Zhang et al. [20] proposed an A3C-based controller. A3C algorithm [21] applies the actor-critic algorithm that runs parallel actor-learners. They trained their model on a simulator and then the trained controller was deployed in a real-life office building. However, their method still suffers from filling the gap between simulation and real environment.

3. DRL-BASED CONTROLLER

As mentioned in [17], the zone temperatures at the next time step are only determined by the current indoor environment conditions, weather disturbances, and supplied air by the AC. It is independent of the previous conditions. Therefore, we can formulate the airflow direction control of ACs as the Markov decision process.

We consider the control of a typical electric refrigerant-based AC unit that is commonly used in households in Japan. They are air sourced and bivalent. As the target application is the AC, we assume that the AC is deployed in residential houses. We target relatively large rooms in the houses (i.e., living rooms) where temperature distribution may variate.

Several different spots in the room are chosen as zones. We assume that those zone air temperatures can be predicted from measurable information such as infrared thermal images by applying techniques like proposed in [22]. The infrared image can be taken by array sensors that are often equipped to ACs. The proposed method aims to control the airflow direction of the ACs to keep a uniform comfort level. ASHRAE standard [23] shows that the comfort temperature is between 19 to 24 ℃. In this study, room temperature is considered as the comfort metrics and the temperature is aimed to be controlled around 22 ℃. Our proposed controller aims to reduce the variance of zone temperatures, $T_{z_0}, \ldots, T_{z_n}$.

3.1 Control framework

RL controllers are trained by agent-environment interaction. As the RL environment, we used the CFD software, Autodesk CFD Ultimate 2019 [24]. The analysis is conducted by the finite element method and uses k-ω turbulence models. We use real-life data for the weather information of outdoor ambient temperature and solar radiation. The outdoor ambient temperature data are extracted from the database of Japan Meteorological

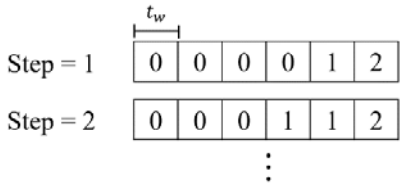


Fig.2 The concept of the proposed action slot.

Agency [25]. The solar radiation is simulated by the CFD software from the location, date, and time information. The AC is modeled as a heat exchanger device.

The CFD simulator communicates with the agent via socket communication using its Python API. The simulator receives control action, a_t from the agent. The agent is implemented at the remote server. It receives the output state s_t from the simulator and trains the DRL model. Fig.1 shows our proposed DRL control framework.

3.2 DRL-based controller design

Control action: The controllable variables are temperature set-point, airflow direction, and airflow volume. For simplicity, the proposed controller only selects the airflow direction. The airflow volume is fixed as constant. The temperature set-point is decided by a separate algorithm; it uses a simple corrective strategy. When the average zone temperature (T_{av}) is smaller than the lower threshold of the room temperature (T_{goal_min}), the temperature set-point (T_{set}) of the next step is increased by 1 ℃ if T_{av_t} is smaller than $T_{av_{t-1}}$. Otherwise, T_{set} is not changed. T_{set} is decided in the same manner for when T_{av_t} is larger than $T_{av_{t-1}}$. T_{set} is controlled between the maximum and minimum threshold, T_{set_max}, and T_{set_min}, respectively. In this research, T_{set_max} and T_{set_min} are 32.0 ℃ and 52.0 ℃, respectively. T_{goal_min} and T_{goal_max}, is set 21.5 and 22.5 ℃, respectively.

Airflow direction is the action space of the proposed RL controller. In this study, we consider an AC which has three controllable fan directions (i.e., 0: left, 1: center, and 2: right). To simulate the realistic behavior, we chose two directions with small airflow volume (V_{small}) and one with a large volume (V_{large}). In this research, we experimentally decided V_{small} and V_{large} as 1 and 10 m^3/min, respectively. We observed different thermal responses of the indoor environment and required control interval; change of the airflow direction does not affect the zone temperature immediately. This slow response of room temperature causes the dilemma that smaller control interval does not show the significance of each action while larger interval changes the temperature too drastically. To solve this problem, we introduce the action slot which contains a set of different directions. Fig. (2) explains the concept of the proposed action slot with six windows. The slot consists of windows with the interval of t_w. For each step, an action slot is selected by the agent. The numbers in each window present the selected direction with V_{large}. The action slot is a combination of different directions. Since large action space slows learning convergence, the action slots are calculated under the following rules:

- Each slot contains all directions of 0, 1, and 2.
- The directions are assigned at the fixed ascending order.

In this research, we used the action slot of 6 windows which is determined experimentally. This gives the action space with the order of 10.

System state: State is the input to the agent. We define that state s_t consists of the information of weather (i.e., time and outdoor ambient temperature), control settings (i.e., airflow direction and temperature set-point), and indoor thermal environment (i.e., zone temperatures and the difference between zone temperatures of $t-1$ and t). Zone temperature is scaled to 0 to 1. We consider two important influencing factors here; weather and time-series information of the indoor thermal environment. For weather information, we used time and outdoor ambient temperature. We used the difference in zone temperatures to consider the time-series behavior of the indoor environment.

Reward function: The agent aims to maximize the accumulative reward R. In each control step, an immediate reward r is given to evaluate the chosen action. We define the reward function as in Eq. (3).

$$r = \begin{cases} 0.1, & T_{z_{max}} - T_{z_{min}} \leq 0.5 \\ -0.1 - \frac{1}{n}\sum_{i=1}^{n}(T_{z_i} - \overline{T_z}), & Otherwise \end{cases} \quad (3)$$

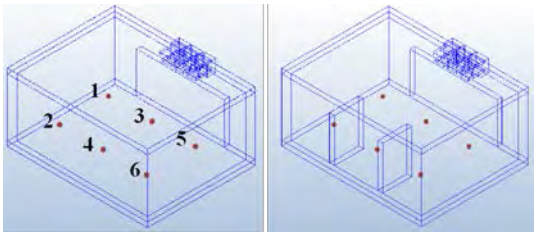
We consider that the difference of 0.5 ℃ between T_z is acceptable and gives a constant positive reward. When the difference between maximum and minimum zone temperature ($T_{z_{max}}$ and $T_{z_{min}}$, respectively) is larger than 0.5 ℃, a penalty that is proportional to the variance of each zone temperatures, $T_{z_1}, \ldots, T_{z_n}$.

Q-network design: The neural network architecture of main and target Q-network proposed in [11] is used; a linear network with 3 layers It has 2 hidden layers; the rectified linear unit (ReLU) is used as the activation function. The number of neurons for the hidden layers is 32. The input to the network is the state s_t. The output is the Q-values for n actions (i.e., $Q(s_t, a_t^1), \ldots, Q(s_t, a_t^n)$).

3.3 Control flow

Our DRL-based AC controller algorithm is presented in Algorithm 1. The agent is trained by N episodes that consist of L steps. In this research, we set N and L as 30 and 90, respectively. Since we use an action slot with 6 windows with a size of 20 s (i.e., 2 min for 1 step), 1 episode is equivalent to 3 h. We consider that this is reasonable because the time of continuous AC usage is considered to be around 3-4 h.

At the beginning of each episode, the environment is reset by the random weather information. To give various states in the purpose of increasing learning performance,



(a) Without partitions (b) With partitions

Fig.3 Simulated environment

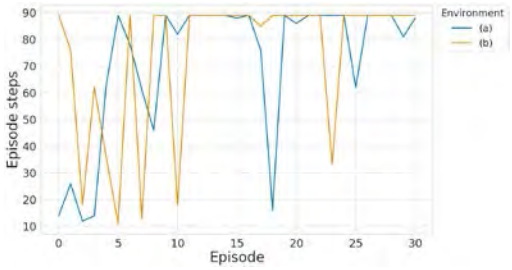


Fig.4 Episode steps obtained for each environment.

the AC is run for Δt_{init} with random actions (lines 3 and 4). This is equivalent to 10 min. Then T_{set} is calculated. A control action is chosen by using the ε-greedy policy (lines 7 and 8). After running the simulation, the obtained transition tuple (s_{t-1}, a, s_t, r_t) is stored in the memory M (line 11). The main Q-network Q_m is trained with the minibatch drawn from M (line 13) by the backpropagation method. The episode loop is finished if the difference between $T_{z_{max}}$ and $T_{z_{min}}$ is larger than T_{diff_max} which is considered to be a failure (lines 14 and 15). In this research, we set T_{diff_max} as 4 K. The target Q-network Q_t is updated by copying Q_m periodically (lines 17 and 18). We updated Q_m every 2 episodes.

Algorithm1 DRL-based AC controller

1: Initialize experience replay memory M, main Q-network Q_m and target Q-network Q_t
2: **for** $episode = 1$ to N **do**
3: Reset building environment to initial state
4: Run simulation for Δt_{init} to initialize s_t
5: **for** $step = 1$ to L **do**
6: Calculate T_{set}
7: $\epsilon = \max(\epsilon - \Delta\epsilon, \epsilon_{min})$
8: $a_t = \begin{cases} argmax_{a'} Q_m(s_t, a'), & probability\ \epsilon \\ random(n), & Otherwise \end{cases}$
9: Run simulation with action a_t
10: Calculate r_t
11: $M \leftarrow (s_{t-1}, a, s_t, r_t)$
12: $minibatch \leftarrow M$
13: Train Q_m with $minibatch$
14: **end if**
15: $T_{z_{max}} - T_{z_{min}} < T_{diff\_max}$
16: **if** $episode\%T = 0$ **then**
17: $Q_t \leftarrow Q_m$
18: **end for**
19: **end for**

Table 1 Parameter settings in the proposed DRL algorithm.

Description	Value
Number of episodes, N	30
Number of steps, L	90
Batch size	128
Learning rate, η	0.0005
Discount rate, γ	0.99
Initial ϵ, ϵ_{init}	0.5
Minimum ϵ, ϵ_{min}	0.05
Decay rate of ϵ, $\Delta\epsilon$	0.008

4. EXPERIMENTAL RESULTS

4.1 Experimental design

In this research, we used a simulated room of size 5.40 m (width) × 4.05 m (length) × 2.40 m (height) as shown in Fig. 3. We conducted the experiment in two types of rooms; (a) without partitions and (b) with two partitions with size 0.01 m (width) × 1.20 m (length) × 1.80 m (height) from 0.16 m from the East and West wall. They have a window in the direction of the South. The wall and the window receive solar radiation. The ceiling and the floor are insulated. Temperatures at all six zones (i.e., $T_{z_1}, \ldots, T_{z_6}$) at 0.1 m from the floor are used as the system state s_t. Temperatures at 4 zones, T_{z_1}, T_{z_2}, T_{z_5}, and T_{z_6} are used to calculate the reward r_t. The AC is modeled as three separate heat exchanger systems. All outlets are at the direction of 30 ° with the ceiling. Left, center, and right outlets are faced at 30 °, 90 °, and 150 ° from the wall, respectively.

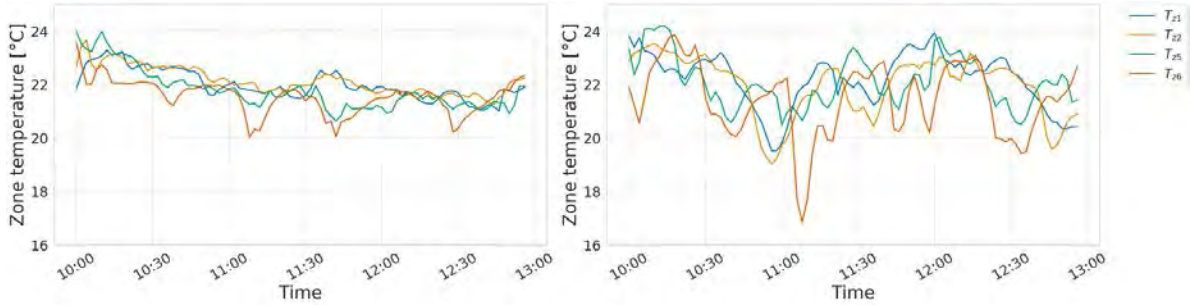
The weather data were split into training and testing datasets. For training, weather data were chosen randomly from the period between 1st to 20th January 2019. For testing, five different dates and time periods were chosen; date of 21st to 25th January 2019, with the starting time of 00:00, 05:00, 10:00, 15:00, and 20:00. Simulation of 3 h period was conducted for evaluation for each test dataset.

The simulator was run on 2.20 GHz Intel(R) Xeon(R) Silver 4114 CPU processor with 96.0 GB memory and running on a Windows 10 operating system. The agent was trained on 2.40 GHz Intel(R) Xeon(R) CPU E5-2680 v4 processor with 132 GB memory and running on an Ubuntu 16.04.5 operating system. The training time was 25.1 h and 29.1 h for (a) and (b), respectively. The parameters in our DRL controller are summarized in Table 1. They are determined suboptimal from pre-experiments.

Since no previous works exist for airflow direction control, the performance of the proposed method is compared with the fixed and simple rule-based methods. The temperature is decided by the same method as the

Table 2 Summary of experimental results.

	Date	21 Jan. 00:00	22 Jan. 05:00	23 Jan. 10:00	24 Jan. 15:00	25 Jan. 20:00	21 Jan. 00:00	22 Jan. 05:00	23 Jan. 10:00	24 Jan. 15:00	25 Jan. 20:00
		Average variance of T_{z_1}, T_{z_2}, T_{z_5}, and T_{z_6} [K^2]					Energy consumption, × 10^3 [W]				
(a)	Proposed	0.56	1.26	0.21	0.64	0.37	3.54	5.67	2.44	2.92	3.88
	Rule-based	1.17	2.06	0.95	1.03	0.89	3.84	6.16	3.13	3.44	3.96
	Fixed	1.36	0.64	1.19	1.28	0.83	5.75	6.81	4.55	4.84	5.99
(b)	Proposed	0.57	1.19	0.53	0.85	0.68	5.38	6.90	3.96	4.30	5.24
	Rule-based	0.74	1.12	0.89	0.80	1.05	4.84	6.67	4.07	4.30	5.18
	Fixed	0.70	0.82	0.71	0.67	0.73	6.53	7.23	5.46	6.02	6.48



(1) Proposed method (2) Rule-based method

Fig. 5 Zone temperatures for 23 Jan. 2019, (a).

proposed method for both. For the fixed method, only the temperature is controlled and the airflow direction is not changed; the airflow volume is always V_{large} at the center and V_{small} at the left and right. For the rule-based method, the airflow direction is decided from the average temperature of each zone; three sets of average temperatures, (T_{z_1}, T_{z_2}), (T_{z_3}, T_{z_4}), and (T_{z_5}, T_{z_6}) are calculated as the reference temperatures, T_{r_1}, T_{r_2}, and T_{r_3}, respectively. The action slot is chosen by comparing the difference between T_{r_i} and their average, $T_{r\_diff_i}$ and the ratio of the period of each direction. The ratio of the maximum and each $T_{r\_diff_i}$ is calculated as Eq. (4) and converted to the ratio. The action slot with the closest ratio of the fan direction is chosen.

$$\frac{\max\left(T_{r_{diff_i}}\right) - T_{r_{diff_i}} + \max\left(T_{r_{diff_i}}\right)}{\max\left(T_{r_{diff_i}}\right)} \quad (4)$$

4.2 Experimental results

Episode steps during training: Fig. 4 shows the obtained episode steps during the training. The controller was able to achieve the maximum step of 90 more frequently after around episode 10 for both (a) and (b). The number of obtained steps sometimes variates after episode 10. This may be because of random actions that are chosen by the ε-greedy policy. The random action can increase the variance of zone temperatures drastically. For (b), the obtained episode steps variates more than (a) before episode 10. This is because temperature difference tends to be larger for (b) than (a) due to the partitions. Our proposed controller successfully learns the optimized airflow directions and the maximum episode steps converge similarly to the case of (a).

Performance of the proposed DRL controller to obtain uniform comfort: The average variance of zone temperatures T_{z_1}, T_{z_2}, T_{z_5}, and T_{z_6} for each step is calculated to compare if the room was kept under the uniform comfort for the proposed, rule-based, and fixed methods. Table 2 summarizes the experimental results. For (a), the average variance of T_{z_1}, T_{z_2}, T_{z_5}, and T_{z_6} is smaller for the proposed method than the rule-based and fixed method except for 22nd January. We observed a low outdoor temperature for 22nd January. When the temperature difference between indoor and outside is large, the indoor temperature decreases faster for the fixed method. Then the fixed method increases the temperature set-point; warmer air diffuses faster and achieves uniform temperature than the proposed method. However, the proposed method improves the average variance of zone temperatures by 53.0 and 46.9 % for the average than the rule-based and fixed method, respectively. The rule-based method may have failed to achieve uniform comfort because of missing the ability to predict the behavior of air mixing. Since our proposed method uses the DRL technique, it effectively selects the airflow direction to maximize the future accumulated reward.

Fig. 5 illustrates the behavior of zone temperatures for 10:00-13:00 23rd January 2019, (a). A relatively more uniform temperature is obtained for the proposed method compared to the rule-based method. The maximum difference between zone temperatures is 2.48 ℃ as the worst around 11:40. After a large variance of zone temperatures is observed, the airflow direction is changed to cancel the temperature difference. Although the rule-based method is successful to cancel the temperature difference after the variance becomes larger, the

Table 3 Average variance of T_{z_1}, T_{z_2}, T_{z_3}, T_{z_4}, T_{z_5}, and T_{z_6} for (b) [K^2]

Date	21 Jan. 00:00	22 Jan. 05:00	23 Jan. 10:00	24 Jan. 15:00	25 Jan. 20:00
Proposed	0.72	1.58	0.64	0.84	0.75
Fixed	1.25	1.51	1.27	1.51	1.32

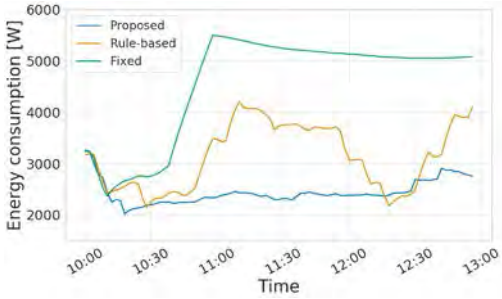


Fig. 6 Energy consumption for 23 Jan. 2019, (a).

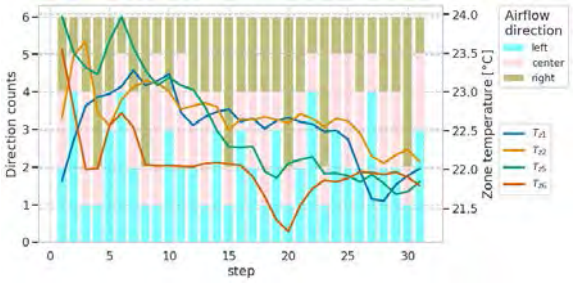


Fig. 7 Selected directions by the proposed method for 23 Jan. 2019.

temperature change is more dynamic than the control of the proposed method which can cause thermal discomfort.

For (b), the average variance of T_{z_1}, T_{z_2}, T_{z_5}, and T_{z_6} is smaller on 21st, 23rd, and 25th January for the proposed method than the rule-based method. For other days, the performance of both methods was similar. In the case of (b), the existence of two partitions reduces the heat exchange between zones and deciding the airflow direction depending on each zone temperature is considered to be effective to obtain the uniform temperature within the room. Our proposed method successfully learns the behavior to decide the airflow direction from the zone temperatures.

However, the fixed method outperforms the proposed method for 22nd and 24th January. Two reasons are considered; the higher temperature set-point obtained for fixed method and the symmetry placement of the partitions. Because of the partitions, $T_{z_1}, T_{z_2}, T_{z_5}$ and T_{z_6} are likely to get cooled to decrease the average room temperature and the temperature set-point is increased. Warmer air diffuses faster to decrease the temperature difference, especially for 22nd January when the outdoor temperature is low. Please note that only $T_{z_1}, T_{z_2}, T_{z_5}$ and T_{z_6} are used to calculate the reward. This is done because T_{z_3} and T_4 are close to the AC and likely to get heated more than other zones. In addition, because of the symmetry placement of partitions, (T_{z_1}, T_{z_2}) and (T_{z_5}, T_{z_6}) are likely to be similar. Table 3 shows the average variance of the temperatures of all 6 zones. The variance of 6 zone temperatures is larger for the fixed method than the proposed method except for 22nd January. This explains that T_{z_3} and T_{z_4} is heated much more than other zones. Our proposed method is effective to achieve the uniform temperature within a room even though there are partitions.

Energy consumption: To evaluate the energy performance of our proposed method, the energy consumption of the modeled AC is calculated using Eq. (5) where P is energy consumption, $\dot{m}$ is volume mass of airflow, C_p is specific heat capacity, T_{out} is outlet temperature, and T_{in} is inlet temperature to the AC.

$$P = \dot{m} C_p (T_{out} - T_{in}) \quad (5)$$

We consider T_{out} as the temperature set-point and T_{in} as the average of zone temperatures. Please note that the AC is modeled as a heat exchanger in the CFD software. The energy consumption of a real AC considers the coefficient of performance (COP); because of efficiency, the energy consumption of a real AC is much smaller than the values indicated in the result. From Table 2, the energy consumption is smaller for the proposed method than the fixed and rule-based method for both (a) and (b). The reduction is 57.6 % as the best and 4.56 % as the worst. The result suggests that keeping the room temperature uniformly can also save energy consumption. This can be caused by reducing the temperature difference between room and outdoor temperature in certain zones. If some zones are heated than other zones, the heat loss would be larger from the zones which may decrease the total energy efficiency. Fig. 6 illustrates the energy consumption for 23rd January 2019, (a). For the fixed method, energy consumption is the largest of the three methods. The change in energy consumption is more moderate for the proposed method than the rule-based method. For the rule-based method, it suggests that the uneven temperature distribution around 11:00 triggers a sudden increase in temperature set-point and increases energy consumption. The room is heated enough to decrease the temperature set-point around 12:00. However, the room temperature is cooled too much and the temperature set-point increases again after 12:30. On the contrary, our proposed method keeps the energy consumption around 2000-2500 W from 10:30-12:30 and then moderately increases its energy consumption. It implies that the change in temperature set-point is also moderate. Not only achieving low energy costs, but it can also provide pleasant temperature change inside a room by changing the temperature set-point moderately.

Selected direction of proposed DRL controller: We provide a deeper analysis of the behavior of our DRL method. Fig. 7 shows the first 30 steps (i.e., 10:00-11:00) of the selected airflow direction by the proposed method for 23rd January 2019. The left y-axis shows the direction counts within the chosen action slot and the right y-axis shows the output temperature of the step. We can observe

that the flexible directions are selected depending on each zone temperature. For example, at step = 0, T_{z_1} and T_{z_2} are smaller than T_{z_5} and T_{z_6} and at step = 1, the action slot with a larger count for left direction is chosen. On the contrary, when T_{z_5} and T_{z_6} are smaller than T_{z_1} and T_{z_2} at step = 19, action slot with larger count for right direction is chosen at step = 20 which successfully increases T_{z_5} and T_{z_6}. These direction selections are reasonable that confirms the performance of our proposed DRL controller.

5. CONCLUSION

We proposed a DRL-based AC controller that achieves a uniform comfort within a room while reducing the energy consumption by optimizing the airflow direction. The proposed method is evaluated against fixed and rule-based methods. The proposed DRL-based controller shows the improvement in the average variance of zone temperatures by 21.3 % and reduction in energy consumption by 35.0 % in the average. Our experimental results show that the proposed method makes a moderate temperature change in the indoor environment. Furthermore, the proposed DRL controller can be applied to the houses with partitions.

The possible future direction of our research is incorporating zone temperature prediction techniques. Infrared images which are obtained by array sensors can be used to predict zone temperatures.

ACKNOWLEDGEMENT

This work was supported by JST CREST Grant Number JPMJCR19K1, MEXT/JSPS KAKENHI Grant (B) Number JP16H04455 and JP17H01739. The authors would like to thank Panasonic Corporation, Appliances Company for providing advice for this research.

REFERENCES

[1] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy and Buildings*, Vol. 40, No. 3, pp. 394–398, 2008.

[2] Y. Agarwal, B. Balaji, S. Dutta, R. K. Gupta, and T. Weng, "Duty-cycling buildings aggressively: The next frontier in HVAC control," In *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks, IPSN'11*, pp. 246–257, 2011.

[3] A. Ghahramani, F. Jazizadeh, B. Becerik-Gerber, and S. Astani, "A knowledge based approach for selecting energy-aware and comfort-driven HVAC temperature set points," *Energy and Buildings*, Vol. 85, pp. 536–548, 2014.

[4] M. Abedi, F. Jazizadeh, B. Huang, and F. Battaglia, "Smart HVAC Systems - Adjustable Airflow Direction," In *Advanced Computing Strategies for Engineering*, pp. 193–209, 2018.

[5] A. Afram and F. Janabi-Sharifi, "Theory and applications of HVAC control systems–A review of model predictive control (MPC)", B*uilding and Environment*, Vol.72, pp.343-355, 2014.

[6] A. Afram and F. Janabi-Sharifi, "Gray-box modeling and validation of residential HVAC system for control system design," *Applied Energy*, Vol. 137, pp. 134–150, 2015.

[7] W. J. Cole, K. M. Powell, E. T. Hale, and T. F. Edgar, "Reduced-order residential home modeling for model predictive control," *Energy and Buildings*, Vol. 74, pp. 69–77, 2014.

[8] E. Barrett and S. Linder, "Autonomous HVAC control, a reinforcement learning approach", In *ECMLPKDD'15 Proceedings of the 2015th European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 3–19, 2015.

[9] C. J. C. H. Watkins and P. Dayan, "Q-learning", *Machine Learning*, Vol. 8, No. 3–4, pp. 279–292, 1992.

[10] B. Li and L. Xia, "A multi-grid reinforcement learning method for energy conservation and comfort of HVAC in buildings," *IEEE Int. Conf. Autom. Sci. Eng.*, vol. 2015-Octob, pp. 444–449, 2015.

[11] V. Mnih *et al.*, "Playing Atari with Deep Reinforcement Learning", http://arxiv.org/abs/1312.5602.

[12] V. Mnih *et al.*, "Playing Atari with Deep Reinforcement Learning", http://arxiv.org/abs/1312.5602, 2013.

[13] D. Nikovski, J. Xu, and M. Monaka, "A Method for Computing Optimal Set-Point Schedule for HVAC Systems", In *Proceedings of the 11th REHVA World Congress CLIMA*. 2013.

[14] A. Zenger, J. Schmidt, and M. Krödel, "Towards the Intelligent Home: Using Reinforcement-Learning for Optimal Heating Control", In *Annual Conference on Artificial Intelligence,* pp.304-307, 2013.

[15] A. Nagy, H. Kazmi, F. Cheaib, and J. Driesen, "Deep Reinforcement Learning for Optimal Control of Space Heating", In *Proceedings of BSO 2018: 4th Building Simulation and Optimization Conference*, pp. 96-103, 2018. .

[16] L. Yang, Z. Nagy, P. Goffin, and A. Schlueter, "Reinforcement learning for optimal control of low exergy buildings", *Applied Energy*, Vol. 156, pp. 577–586, 2015.

[17] T. Wei, Y. Wang, and Q. Zhu, "Deep Reinforcement Learning for Building HVAC Control", In *Proceedings of the 54th Annual Design Automation Conference,* pp. 1–6, 2017.

[18] G. Gao, J. Li, and Y. Wen, "Energy-Efficient Thermal Comfort Control in Smart Buildings via Deep Reinforcement Learning", http://arxiv.org/abs/1901.04693, 2019.

[19] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning", in *Proceesings of 4th International Conference on Learning Representations, ICLR 2016*, 2016.

[20] Z. Zhang and K. P. Lam, "Practical implementation

and evaluation of deep reinforcement learning control for a radiant heating system", In *Proceedings of the 5th Conference on Systems for Built Environments*, pp. 148–157, 2018.
[21] V. Mnih *et al.*, "Asynchronous Methods for Deep Reinforcement Learning", http://arxiv.org/abs/1602.01783, 2016.
[22] C. Porras-Amores, F. R. Mazarrón, and I. Cañas, "Using quantitative infrared thermography to determine indoor air temperature", *Energy and Buildings*, Vol. 65, pp. 292-298, 2013.
[23] "ASHRAE,. Standard 55:2004. Thermal environment conditions for human occupancy," Atlanta, GA, USA, 2004.
[24] "Autodesk CFD Ultimate 2019.", https://www.autodesk.com/products/cfd/overview.
[25] "Japan Meteorological Agency.", https://www.jma.go.jp/jma/index.html.