Comfort and energy management of multi-zone HVAC system based on Multi-Agent Deep Reinforcement Learning

Wanlin Bo
College of Control Science and Engineering
Shandong University
Jinan, China
Email: 2081502706@qq.com

Xinli Wang
College of Control Science and Engineering
Shandong University
Jinan, China
Email: wangxinli@sdu.edu.cn

Gang Jing
School of Transportation and Logistics Engineering
Shandong Jiaotong University
Jinan, China
Email: 365173573@qq.com

Haojin Xu
College of Control Science and Engineering
Shandong University
Jinan, China
Email: 2446612062@qq.com

Lei Jia
College of Control Science and Engineering
Shandong University
Jinan, China
Email: jialei@sdu.edu.cn

2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA) | 979-8-3503-1220-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIEA58696.2023.10241916

***Abstract*—Energy consumption generated by HVAC systems accounts for more than 40% of total energy usage in commercial buildings. There are many studies focusing on improving comfort conditions for occupants in multi-zones while reducing energy consumption. For example, the widely used Multi-agent Deep Deterministic Policy Gradients algorithm, however, the principle of centralized training and overestimation strategy, which will make the algorithm fall into a sub-optimal process and have the slowly training speed. In order to solve this problem, we decided to adopt multi-agent twin-delayed deep deterministic policy gradient algorithm for the HVAC control of multi zone building. In this work, we assume that one zone is controlled by its corresponding agent and consider the control problem as a Markov game first. Then we adopt delayed update strategy and design a target policy smoothing technique for our proposed method. Furthermore, two value functions will be designed for each agent to prevent overestimation. We validate the effectiveness of the method for HVAC systems by simulation using TRNSYS software.**

Keywords*—*HVAC systems, multi-agent deep reinforcement learning (DRL), energy cost, comfort control

I. Introduction

Heating, Ventilation, and Air Conditioning (HVAC) systems in buildings are responsible for over 35% of carbon emissions. Building HVAC energy saving and low-carbon operation is one of the effective measures to relieve energy shortage and improve building energy efficiency, as well as an energy management strategy to effectively reduce peak load and stabilize the whole system operation[1]. Meanwhile, it is very necessary to keep a comfortable environment to guarantee health and productivity of occupants in the building. Thereby, we need to adopt a HVAC control approach to achieve a balance between comfort and energy consumption.

There are many researches about improving HVAC control method to minimize the energy consumption while ensuring thermal comfort for occupants of the zones[2]. For example, proportional integral-derivative (PID)[3]、model predictive control algorithm (MPC)[4] and Lyapunov optimization techniques[5]. These methods are all model-based and need detailed building thermal dynamics models. However, it is not easy to model the detailed and accurate building thermal dynamics for HVAC control. Besides, it is difficult to apply a generalized model for all building due to the differences of environments.

With the development of artificial intelligence, data-driven methods based on artificial neural network[6]-[8] are adopted to figure out this problem gradually, such as deep reinforcement learning[9], which will learn the optimal control strategy through continuous interactive training with the building environment without a thermal dynamic model. For example, in[10], the authors proposed the method of Deep Q-Network (DQN) in order to keep the indoor temperature within the appropriate range while saving energy cost as more as possible. Deep Deterministic Policy Gradients (DDPG) method is adopted to reduce energy consumption while guaranteeing the thermal comfort requirements for occupants in single zone[11]. Based on Dueling Q-network architecture, Ding et al. proposed a method for the control of four building subsystems[12]. In[13], DDPG method is used to optimize the continuous thermal control strategy of multi-zone HVAC system by changing the setpoint in residential building. However, when single agent controls the temperature of multiple zones through HVAC system, the training dimension of neural network will increase observably as the number of zones increases. The training speed is slowly and the only agent can not quickly identify regional differences. Therefore, the MADRL control method is adopted gradually. In[14], multi-agent deep reinforcement learning control algorithm is proposed to save cost and maintain the indoor temperature for multi-zone buildings. There are also optimizing the multizone thermal control using Multi-agent Deep Deterministic Policy Gradients[15].

In the training process of multi-agent reinforcement learning, according to the principle of centralized training and over-estimation strategy[16], which will make the algorithm fall into a sub-optimal process, leading to a long training time and unable to rapidly converge to a stable state. This may cause significant time cost problems, especially in large building simulations. Therefore, we propose the method based on multi-agent twin-delayed deep deterministic policy gradient algorithm for the HVAC control of the multi-zone building, each zone is controlled by its corresponding agent. All agents can learn from each other and refer to others' information to accelerate the learning process of control strategy. Meanwhile, this work intends to realize the balance between comfort and low energy cost under dynamic price with the consideration of random zone occupation without the detailed building dynamic model.

II. MODEL FORMULATION

In this section, the HVAC control model and basic principle are described firstly. Next, the control problem of multi-zone HVAC will be described as a Markov decision process and then we introduce essential elements of MDP.

A. HVAC Control Model

We assume that there is a commercial building composed of N different zones and the indoor temperature and humidity would be changed through the HVAC control system in these zones[14]. There are an air handling unit (AHU) for the whole building in HVAC system and each zone configures a variable air volume (VAV) box individually, the details are shown in Fig. 1[17]. Here, we consider the air flow rate as the control input for each zone, which is controlled by the fan and VAV box of each zone[18]. In order to ensure generalization, we consider the case of cooling control for all zones in this work.

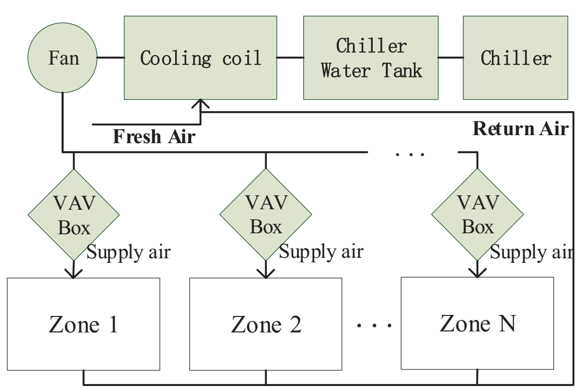

Fig. 1. HVAC control architecture for *N*-zone.

B. Reinforcement Learning

The indoor thermal condition of the present moment in the room is only connected with the state parameters of the previous moment[19]. Therefore, it's feasible to formulate the control problem of multi-zone HVAC as a Markov decision process (MDP) and use RL method to solve it, which brings the problem into the scope of the Bellman Optimality Equation. Recursive Bellman Equation is

$$V(s) = \max_a [R(s,a) + \gamma V(s')] \tag{1}$$

Where s is the current state, s' is the next state having taken action a and get the corresponding reward $R(s, a)$, $V(s)$ is the current state value, γ is the discount rate, $V(s')$ is the next state value. There are four essential elements in MDP and they are state, action, state transition probability, and the corresponding reward. We assume N zones in the HVAC control problem environment, as shown in Fig. 2. More specific details are defined as follows:

1) State

The state is the information that agents obtain from the environment and is a set of factors which can express current indoor and outdoor condition. We consider N zones in the building, zone i is managed by agent i for $0 \le i \le N-1$. The environmental state of the building is the indoor and outdoor condition (temperature and humidity) monitored by sensors at the beginning of each time period, which greatly affects the comfort of occupants. In addition, the zones occupancy O_t at time slot t should also be considered that influences agents' decision. What's more, all agents aim to minimize HVAC energy cost, which is related to Time-varying electricity price λ_t. Considering all states of the N zones, it can be expressed as $s_t = (s_t^0, s_t^1, \cdots, s_t^{N-1})$. According to the actions taken by the N agents at time slot t, the environment moves into the next state at time slot t+1.

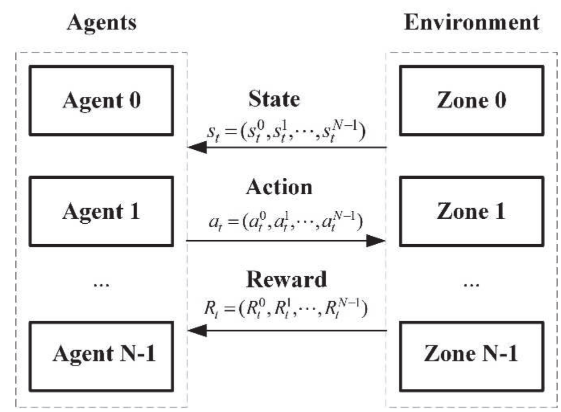

Fig. 2. MDP control architecture of multi-agent.

2) Action

In zone i ($0 \le i \le N-1$), agent i will take actions to adjust the indoor environment state of zone i according to the local observation. Here, it refers to the air flow rate that controlled by the fan control signal of each zone at time slot t for each zone. The action of agent i is determined using control strategy μ^i as $a_t^i = \mu_t^i(s_t^i)$ .Then we have actions $a_t = (a_t^0, a_t^1, \cdots, a_t^{N-1})$.

3) Reward

Reward is the feedback of the environment to the action, which plays a crucial role for the control strategy learning of DRL and influences the convergence process of model training. According to Markov decision theory, the transition from environmental state s_{t-1} to s_t is converted by the action a_{t-1}, and then get the returned reward function r_t. Here, we consider both energy cost and thermal comfort when constructing the reward function. Firstly, energy consumption cost generated by each zone is the energy consumption multiplied by the electricity price. Next, thermal comfort value is a key factor in assessing the satisfaction of occupants with the building's surrounding environment.

In this paper, we adopt ASHRAE 55 and ISO 7730 standards. The predictive mean voting index (PMV) was defined to predict the average vote of occupants about the thermal sensation scale that ranges from −3 to +3, which −3 represents the very cold condition, 0 is neutral and +3 represents extremely hot by Fanger[20]. In applications, PMV is affected by environmental factors, including air temperature T_t^i, relative humidity H_t^i, air velocity V_t^i and average radiation temperature M_t^i, and human factors, including metabolic rate B_t^i and clothing insulation I_t^i. Therefore, thermal comfort value P_t^i can be expressed as:

$$p_t^i = PMV(T_t^i, H_t^i, V_t^i, M_t^i, I_t^i, B_t^i) \tag{2}$$

according to the above analysis, the reward function R is defined that considering both the energy cost and the thermal comfort values of occupants as follows:

$$r_t^i = R(e_t^i \lambda_t, p_t^i) = -\alpha \lambda_t e_t^i - \beta p_t^i \tag{3}$$

We formulate a multi-objective optimization function for solving the multizone HVAC control problem and add weight factors to the objective function for different parts.

$$R_t^i = \sum_{j=1}^{T} \gamma_i^{j-1} r_{t+j}^i \tag{4}$$

Where T represent the total optimization range and $\gamma_i \in [0,1]$ is the discount factor that is related to agent i. Moreover, the reward of time slot $t+j$ is r_{t+j}^i and $j = 1,2,\cdots,T$ The optimization objective is the overall discount reward, as follows:

$$J = \max_{a_t^i} \sum_{i=0}^{N-1} \mathrm{E}[R_t^i] = \max_{a_t^i} \sum_{i=0}^{N-1} \sum_{j=1}^{T} \gamma_i^{j-1} (-\alpha \lambda_t e_t^i - \beta p_t^i) \tag{5}$$

It is so challenging to get the state transition probability accurately, so that it is difficult to work out the optimization problem of J. We decided to use the learning-based approach for optimal control strategies learning based on this.

III. SOLUTION ARCHITECTURE

Here, we adopt multi-agent reinforcement learning to solve it, each zone is controlled by its corresponding agent and the agents can update via the shared global information and learn from each other at the same time.

A. Centralized training and Decentralized Execution

We proposed multi-agent twin delayed deep deterministic policy gradient (MA-TD3) algorithm based on the classic actor-critic architecture in order to optimize the multiple zones control policies. Besides, we design the structure of a centralized critic network and decentralized actor network for each agent. During the training, the agents need to share their experience with each other, that is, the critic network of each agent not only needs to consider the local observations that are obtained in the zone and the reaction actions made according to the actor network, but also needs to consider the status information received by all the agents, which is called centralized training. Each agent only relies on the local state when performing an action according to the state, without considering the information of other agents, that is, distributed execution. The specific algorithm is as follows:

B. The proposed method

The DDPG algorithm widely used in the field of HVAC control is the representative of actor-critic methods and in actor-critic framework the policy is updated according to the value estimates of the approximate critic. Function approximation error leads to high estimation and suboptimal strategy, then converge slowly. we select to use the following method to solve these problems.

The actor-critic reinforcement learning algorithm will select the maximum value of the corresponding objective function in the training and updating process, resulting in overestimation deviation. Here, we design six networks for each agent, an actor network and two critic networks and their target networks for updating. For agent i, its critic networks parameterized using $\theta_t^{Q_1}$ and $\theta_t^{Q_2}$, the actor network learn the control policy according to the local state s_t^i and the network parameterized by θ_t^i.

$$a_t^i = \mu_t^i(s_t^i \mid \theta_t^\mu) + N_t \tag{6}$$

$\mu(s_t^i \mid \theta_t^\mu)$ is the decision policy of agent i and the actor network parameterized by θ_t^μ. OU noise N_t is added for better generalization and exploration.

After agent i receives the current observation s_t^i and executes the action a_t^i based on the observation, then agent receives the corresponding reward r_t^i, and the next observation s_{t+1}^i will be observed. The transitions (s_t, a_t, r_t, s_{t+1}) store in the replay buffer B for the critic networks and the actor network training. We will sample a random mini-batch from the replay buffer B for each training. Through minimizing the loss function across all samples as follows, the weight of each agent's critic network will be updated:

$$L_{Q_1} = \frac{1}{K} \sum_{\delta=1}^{K} [Q_{i,1}(s_t^i, a_t^i \mid \theta_{i,1}^Q) - y_{i,1}] \tag{7}$$

$$L_{Q_2} = \frac{1}{K} \sum_{\delta=1}^{K} [Q_{i,2}(s_t^i, a_t^i \mid \theta_{i,2}^Q) - y_{i,2}] \tag{8}$$

To begin, the target networks are used as the deep function approximator to achieve better convergence. Besides, the agents adopt the concept of double Q-learning that calculate the next state value y_i through two target critic networks. Moreover, agent i will select the minimum of two values $y_{i,k}$

(k=*1, 2*) in order to reduce the overestimation of Q-value and then we should put the Q-value into the Bellman equation for the calculation of the loss function.

$$\begin{gathered} y_{i,k} = r_t^i + \gamma Q_{i,k}(s_{t+1}^i, a_{t+1}^i \mid \theta_{i,k}^Q + \varepsilon) \\ \varepsilon \sim clip(N(0,\sigma), -c, c) \end{gathered} \tag{9}$$

$$y = r_t^i + \min_{k=1,2} Q_{i,k}(s_{t+1}^i, a_{t+1}^i \mid \theta_{i,k}^Q) \tag{10}$$

Finally, we add the noise to the target actor networks in the process of updating so that can compensate the influence of the high Q-value estimation. And in order to reduce the dependence of actor network update on critic network, we adopt the method of delayed update, in other words, after the critic network updates for many times, the actor network is updated. Actor network is updated according to the following policy gradient:

$$\nabla_{\theta_t^i} J(\theta_t^i) = \nabla_{\theta_t^i} \mu_t^i(s_t^i \mid a_t^i) \nabla_{a_t^i} Q_i(s_t^i, a_t^i \mid \theta_i^Q)|_{a_t^i=\mu_t^i(s_t^i|\theta_t^\mu)} \tag{11}$$

The target networks of agent i should be softly updated as:

$$\begin{gathered} \theta_i^{\mu'} \leftarrow \tau_a \theta_i^\mu + (1-\tau_a)\theta_i^{\mu'} \\ \theta_{i,k}^{Q'} \leftarrow \tau_c \theta_{i,k}^Q + (1-\tau_c)\theta_{i,k}^{Q'} \end{gathered} \tag{12}$$

The specific training process is illustrated in Algorithm 1.

Algorithm 1: MA-DRL Algorithm for multi-zone HVAC control

```
initialize the replay buffer B, number of episodes M and
  the number of time slots L
for agent i = 0,1, ..., N-1 do
    initialize the parameters of the critic network
      Q_{i,k}(s_t^i, a_t^i | θ_{i,k}^Q) and the actor network μ_t^i(s_t^i | θ_t^μ)
    initialize the target networks Q_{i,k}(s_t^i, a_t^i | θ_{i,k}^Q) and μ_t^i(s_t^i | θ_t^μ)
      with θ_{i,k}^{Q'} and θ_t^{μ'}
end for
for episode = 1,2, …, M do
    initialize a random OU noise N_t for action exploration
    all agents observe the initial state of the multi-zone
      environment
    for t = 1, 2, …, L do
        agent i observe the local state s_t^i and select action
          a_t^i = μ_t^i(s_t^i | θ_t^μ) + N_t
        calculate the reward r_t^i and observe the next state s_{t+1}^i
        store transition (s_t, a_t, r_t, s_{t+1}) in the replay buffer B
    end for
    for agent i = 0,1, ..., N-1 do
        sample a random minibatch of K transitions from
          the replay buffer B
        set the target value according to (9) and (10)
        update the critic networks by minimizing the loss
          function (7) and (8)
        if t mod D:
            update the actor network according to (11)
        end if
    end for
    update the target networks for each agent i via (12)
end for
end for
```

IV. Numerical Evaluation

In this section, we will introduce the used simulation method firstly and then evaluate the effectiveness of our proposed approach based on the experimental results.

A. Simulation method

In order to analyze the effectiveness of our approach, we create a HVAC simulation environment in the multizone building through using TRNSYS software[15]. We can obtain the real-time state from TRNSYS and write into MYSQL database through the MATLAB interface and the algorithm learning process will be accomplished in PyTorch[11]. Here, we create four zones in a commercial building HVAC model which is 256 m^2 using Singapore weather and we use time-varying electricity price that refer to local governments' policy at the same time. Each training episode corresponds to one day, with a time slot interval of 15 minutes and there are 96 time slots in each episode.

B. Comparison of the proposed method with the baseline

In this work, the well-trained agents' learning performance will be verified and tested. We design a baseline case in the absence of the RL agents to provide a basis for comparison. The baseline case is fixed actions which focus on minimizing the energy cost and the comfort value of occupants must be kept within 1 at the same time, the actions are always at the suitable value to avoid more energy cost. Random seven days are chosen in June to test our well-trained agents to ensure the generalization. The simulation results are shown as follows, thermal comfort values using our proposed method in the four zones is shown in Fig. 3. Fig. 4 illustrates the thermal comfort values of four zones under baseline that prioritizes energy cost reduction.

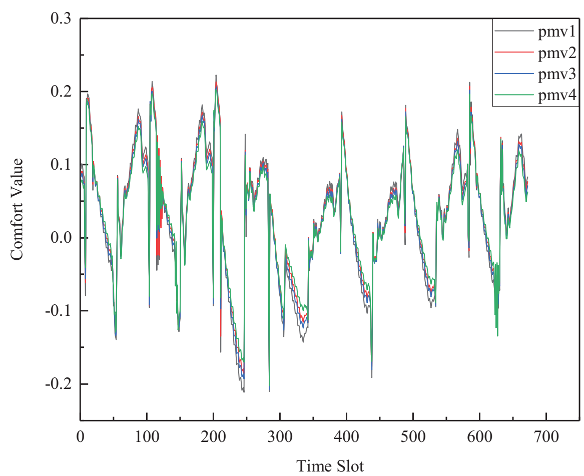

Fig. 3. Thermal comfort values for the zones using the proposed method.

The above results indicate that our method can identify zones' differences after training and adapt to dynamic price. The comfort value of each zone fluctuates within a small acceptable

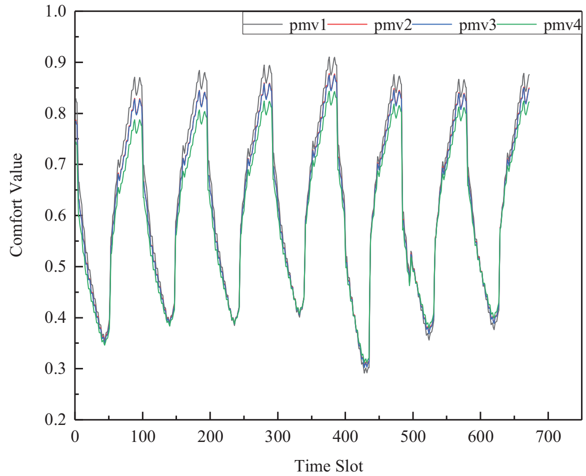

Fig. 4. Thermal comfort values for the zones under baseline method.

range. In contrast, occupants' experience and the comfort values under the baseline that focusing on the energy cost fluctuate greatly as shown in Fig. 4, the baseline method can't identify the difference of the zones. Additionally, the comfort deviation of the four zones in the proposed method is decreased significantly compared with the baseline as shown in the above figures. Meanwhile, the total energy cost of the proposed approach is only 6.47% higher than the baseline as shown in Fig. 5. Therefore, the method can solve the multizone thermal control optimization problem greatly in this work.

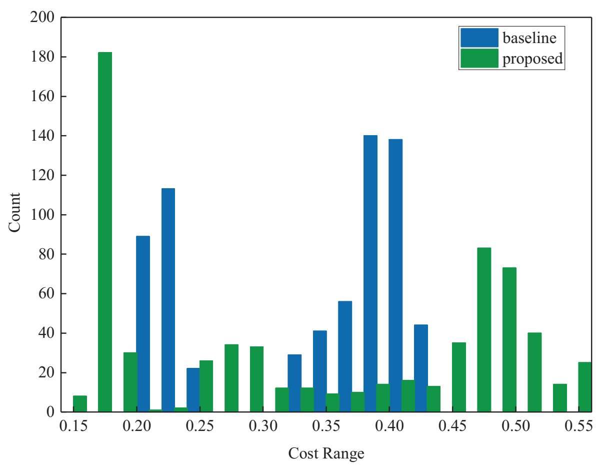

Fig. 5. Comparison of the two methods about energy cost.

C. Comparison of the proposed method with MADDPG

In this part, we analyze the convergence and stability of the proposed method during the training process. The total rewards of four zones in different training episodes is shown in Fig. 6 and then we compare the stability and convergence speed of the training process with two methods. The learning algorithm will store the transition in the replay buffer in 30 episodes, then starts to sample for training. As shown in the figure, the proposed method gradually begins to converge at episode 170, then fluctuates and remains relatively stable, MADDPG converges later by contrast. Besides, the amplitude of the proposed method is smaller and more stable in the all training. What's more, the average rewards of two methods are shown in Fig. 7, the cumulative reward of the proposed method is higher than MADDPG, which represent the more outstanding and effective learning ability.

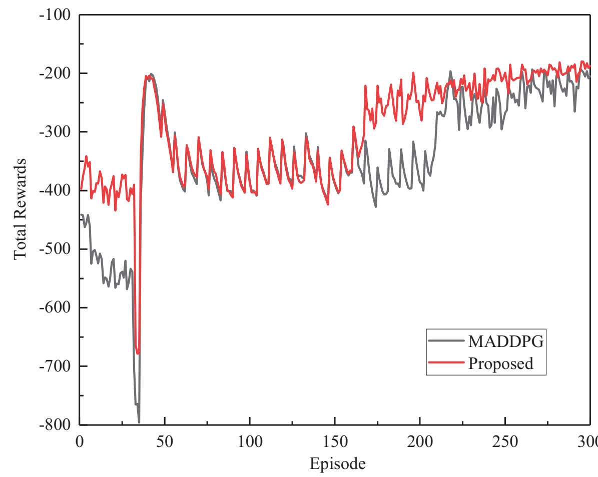

Fig. 6. Training total reward varies with episode of two methods.

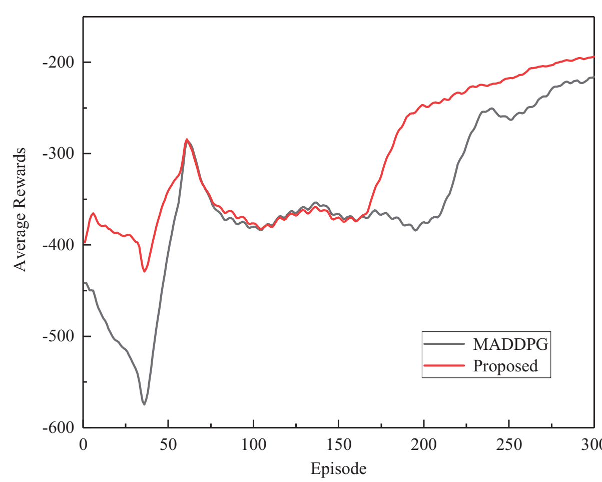

Fig. 7. Training average reward varies with episode of two methods.

V. CONCLUSION

This paper proposed an improved multi-agent deep reinforcement learning method for multi-zone HVAC comfort control and energy cost saving. Here we aim to solve the problem in MADDPG that falls into sub-optimal process, unable to rapidly converge to a stable state and result in a high exploration cost. Additionally, we also consider dynamic price in the building and then the multi-zone building HVAC simulation environment is created to evaluate the performance of the proposed method using TRNSYS. The simulation results indicate our method can achieve comfortable thermal condition while can save cost as much as possible in multi-zone building and has better convergence and stability than MADDPG in the training.

Acknowledgment

This work was supported by National Natural Science Foundation of China under Grant 62073194, the Shandong Provincial Key Research and Development Program under Grant 2019JZZY020812, Shandong Provincial Natural Science Foundation General Project under Grant No. ZR2020MF075 and Young Scholars Program of Shandong University.

References

[1] X. Kou *et al.*, "A Scalable and Distributed Algorithm for Managing Residential Demand Response Programs Using Alternating Direction Method of Multipliers (ADMM)," *IEEE Transactions on Smart Grid,* vol. 11, no. 6, pp. 4871-4882, 2020.

[2] A. I. Dounis and C. Caraiscos, "Advanced control systems engineering for energy and comfort management in a building environment—A review," *Renewable and Sustainable Energy Reviews,* vol. 13, no. 6, pp. 1246-1261, 2009/08/01/ 2009.

[3] A. I. Dounis, M. Bruant, M. Santamouris, G. Guarracino, and P. Michel, "Comparison of Conventional and Fuzzy Control of Indoor Air Quality in Buildings," Journal of Intelligent and Fuzzy Systems, vol. 4, pp. 131-140, 1996.

[4] G. Mantovani and L. Ferrarini, "Temperature Control of a Commercial Building With Model Predictive Control Techniques," IEEE Transactions on Industrial Electronics, vol. 62, no. 4, pp. 2651-2660, 2015.

[5] L. Yu, D. Xie, C. Huang, T. Jiang, and Y. Zou, "Energy Optimization of HVAC Systems in Commercial Buildings Considering Indoor Air Quality Management," IEEE Transactions on Smart Grid, vol. 10, no. 5, pp. 5103-5113, 2019.

[6] W. Li *et al.*, "An LSTM and ANN Fusion Dynamic Model of a Proton Exchange Membrane Fuel Cell," *IEEE Transactions on Industrial Informatics,* vol. 19, no. 4, pp. 5743-5751, 2023.

[7] X. Wang et al., "Distributed LSTM-GCN Based Spatial-temporal Indoor Temperature Prediction in Multi-zone Buildings," IEEE Transactions on Industrial Informatics, pp. 1-10, 2023.

[8] L. Jiang, X. Wang, W. Li, L. Wang, X. Yin, and L. Jia, "Hybrid Multitask Multi-Information Fusion Deep Learning for Household Short-Term Load Forecasting," IEEE Transactions on Smart Grid, vol. 12, no. 6, pp. 5362-5372, 2021.

[9] L. Yu, S. Qin, M. Zhang, C. Shen, T. Jiang, and X. Guan, "A Review of Deep Reinforcement Learning for Smart Building Energy Management," IEEE Internet of Things Journal, vol. 8, no. 15, pp. 12046-12063, 2021.

[10] Z. Jiang *et al.*, "Building HVAC control with reinforcement learning for reduction of energy cost and demand charge," *Energy and Buildings,* vol. 239, p. 110833, 2021/05/15/ 2021.

[11] G. Gao, J. Li, and Y. Wen, "DeepComfort: Energy-Efficient Thermal Comfort Control in Buildings Via Reinforcement Learning," IEEE Internet of Things Journal, vol. 7, no. 9, pp. 8472-8484, 2020.

[12] X. Ding, W. Du, and A. Cerpa, “OCTOPUS: Deep reinforcement learning for holistic smart building control,” in *Proc. BuildSys*, 2019, pp. 326–335.

[13] Y. Du et al., "Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning," Applied Energy, vol. 281, p. 116117, 2021/01/01/ 2021.

[14] L. Yu et al., "Multi-Agent Deep Reinforcement Learning for HVAC Control in Commercial Buildings," IEEE Transactions on Smart Grid, vol. 12, no. 1, pp. 407-419, 2021.

[15] J. Li, W. Zhang, G. Gao, Y. Wen, G. Jin, and G. Christopoulos, "Toward Intelligent Multizone Thermal Control With Multiagent Deep Reinforcement Learning," IEEE Internet of Things Journal, vol. 8, no. 14, pp. 11150-11162, 2021.

[16] S. Fujimoto, H. v. Hoof, and D. J. A. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," vol. abs/1802.09477, 2018.

[17] H. Hao, C. D. Corbin, K. Kalsi, and R. G. Pratt, "Transactive Control of Commercial Buildings for Demand Response," *IEEE Transactions on Power Systems,* vol. 32, no. 1, pp. 774-783, 2017.

[18] V. Reppa, P. Papadopoulos, M. M. Polycarpou, and C. G. Panayiotou, "A Distributed Architecture for HVAC Sensor Fault Detection and Isolation," *IEEE Transactions on Control Systems Technology,* vol. 23, no. 4, pp. 1323-1337, 2015.

[19] N. Lu, "An Evaluation of the HVAC Load Potential for Providing Load Balancing Service," *IEEE Transactions on Smart Grid,* vol. 3, no. 3, pp. 1263-1270, 2012.

[20] E. W. Shaw, "Thermal Comfort: analysis and applications in environmental engineering, by P. O. Fanger. 244 pp. DANISH TECHNICAL PRESS. Copenhagen, Denmark, 1970. Danish Kr. 76, 50," vol. 92, no. 3, pp. 164-164, 1972.