User Comfort and Energy Efficiency in HVAC Systems by Q-learning

Sajjad Baghaee*, Ilkay Ulusoy†
*†Department of Electrical and Electronics Engineering, Middle East Technical University, Ankara, Turkey
*sajjad@baghaee.com, †ilkay@metu.edu.tr

***Abstract*—This study focuses on applying Q-learning techniques for an HVAC agent where the agent learns to find the optimal sequence of ventilator rate variations to satisfy user comfort and energy efficiency simultaneously. On-Off and Setpoint control methods are investigated besides the proposed control method under different occupant number. The results show the advantage of the proposed Q-learning method to keep the Indoor Air Quality (IAQ), i.e., the indoor CO2 concentration, at the desired level while operating the HVAC efficiently.**

Keywords*—*Reinforcement learning (RL), Q-learning, HVAC, ventilation system, comfort and energy efficiency.

I. Introduction

Nowadays, the human demands for comfort and welfare have been growing exponentially. On the other hand, environmental issues such as energy and greenhouse gas (GHG) emission have drawn more and more attention. Energy consumption of the U.S. residential and commercial building is 39.6% of all consumption including industrial and transportation [1]. Therefore, buildings can play a significant role in energy conservation and GHG emission reductions by using the state of the art technology. For instance, the new generation of heating, ventilation, and air-conditioning (HVAC) systems provide various options to satisfy the demands. However, due to the dynamics of the building thermal complexity and environment heterogeneity, the traditional rule-based methods cannot satisfy the demands, comfort and energy efficiency [2].

For energy efficient operation of the HVAC systems, various studies have been doing. For a cooling system a nonlinear model is developed and a model predictive control scheme for minimizing energy consumption is demonstrated by [3]. To conserve HVAC costs, authors in [4] propose a scheduling strategy based on activity type and weather forecasting. In [5] researchers propose an online energy management algorithm based on the framework of Lyapunov optimization techniques to minimizes the energy cost and thermal discomfort cost in a long-term time horizon for an HVAC system.

To minimize the energy cost of a sart home, a partially observable Markov decision process (POMDP) approach is shown in [7], where it provides a $10\% - 30\%$ saving over the status quo. Authors in [8] propose a learning-based demand response strategy for an HVAC system to minimize energy cost. For model-free control systems when the controller does not have sufficient information about the environment, trial-and-error based techniques such as reinforcement learning [6] is a feasible control policy. Q-learning [9] is one of the popular types of reinforcement learning methods. In [10] authors use Q-learning in the thermal energy storage system for cost saving to shift loads from on-peak to off-peak. In [11] by considering the occupants and the building thermal behavior, the batch Q-learning has been applied to reduce the energy consumption while maintaining comfortable temperatures. In this study, we use Q-learning approach to learn intelligently the user comfort and energy efficient strategy for operating the HVAC system to keep the ventilation, i.e., the indoor CO_2 concentration, at the desired level.

978-1-5386-1501-0/18/$31.00 © 2018 IEEE

II. LEARNING ALGORITHM

Reinforcement learning (RL) algorithms are popular in scenarios where interaction with the environment and corresponding action is required. This class of learning algorithms is based on the feedback from the environment. Markov Decision Process (MDP) model is the base of the reinforcement learning algorithms, where a system is defined by its states and how it travels from one state to another.

Q-learning is one of the popular types of reinforcement learning methods, where an agent, the states 'S' and a set of possible actions 'A' for each state are defined for it. The agent goes from one state to another state if an action '$a \in A$' is accomplished. It obtains a reward at each state by making control decisions, where the ultimate goal is to maximize the total reward. The maximum total reward is obtained by learning which action is optimal for each state. Any feedback from the environment updates the reward values, therefore the algorithm is capable of making optimal control decisions for any variation of the environment. Moreover, it does not require any a prior knowledge of the model of the environment. Because in each episode, the agent learns the environment model by experimenting the environment and receives the reward. Under sufficient training the agent can find the optimal policy to reach one of the goal states with minimum discomfort (Convergence of Q-learning).

The quality of a state-action pair is calculated by considering the rewards and recursive nature of the Q-function (1). This value is calculated and stored in the Q matrix. Indeed the Q matrix is the brain of the agent, which shows the memory of what the agent has learned through the interaction with the environment. For each episode a random initial state S_t is selected, then an action among the possible ones is executed randomly or using a policy by the agent. It is followed by receiving a reward, updating the Q matrix and moving to a new state S', where the old state and the selected action determine the new state. Now S' is set as the old state ($S_t = S'$). This loop will continue until S_t is one of the goal states. The structure of the Q-learning algorithm shows a simple value iteration update. It starts with the old value for Q and makes an update based on the new feedback as follows:

$$Q^{'}(s(t),a(t)) = Q(s(t),a(t)) +$$

$$\alpha.[r(t+1)+\gamma.\,maxQ(s(t+1),a(t))-Q(s(t),a(t))] \quad (1)$$

- $Q^{'}(s(t),a(t))$: New Q by action $a(t)$ in state $s(t)$.
- $Q(s(t),a(t))$: Old Q of action-state pair at t.
- $max(Q(s(t+1),a))$: Maximum Q value for all possible actions a in the next state.
- $r(t+1)$: Reward after action $a(t)$ in state $s(t)$.
- α: Learning rate. $0 < \alpha < 1$
- γ: Discount factor. $0 < \gamma < 1$
- $s(t)$: State at time t.
- $a(t)$: Action that can be taken at time t and state $s(t)$

As it is explained, the Q value is estimated by iterative approximation and the common issue for any iterative algorithm is convergence. The convergence theorem for Q-learning claims, if each state-action pair is visited infinitely many times, then the estimated value of Q will converge to the optimal value [9]. This optimal Q value defines a stationary deterministic optimal policy for decisions. In (1) the learning rate (α) affects the number of iteration to converge an optimal policy, however in this equation, if the discount factor γ is closer to zero, it forces the agent to consider the immediate rewards, while if the γ is selected near to 1, long-term high rewards will be considered.

III. Methodology

The smart-building technology aims to enhance automation and convenience with reduced energy consumption. The majority of amenities in a common building use traditional control methods. For instance, the control unit of part of the available HVAC systems apply On-Off scheduling strategy to regulate any variation such as CO_2 or temperature by maximum operation capacity. While other portions use PI control, which contains the adjustable components to control the operating capacity of the amenity. The inputs for the available HVAC systems are temperature, CO_2, relative humidity, air quality or differential pressure. This is obvious that the mentioned inputs could not satisfy the comfort and energy efficiency simultaneously. Therefore, providing the control unit with other inputs such as occupant number and user preference and applying smart strategies like learning methods can meet the comfort and energy efficiency demands.

One of the parameters that defines the Indoor Air Quality (IAQ) is CO_2 concentration. In this study, we consider an HVAC system for providing fresh outdoor air to dilute interior CO_2 contaminants, which is generated by the occupants. The acceptable range for indoor CO_2 concentration is between 600 ppm to 1000 ppm, where for a good indoor air quality 800 ppm can be set as a logical setpoint. Therefore, in this study, the HVAC provides an amount of fresh air from outdoor to the indoor that is sufficient to maintain CO_2 concentration at an acceptable range. Furthermore, RL has applied to control the smart HVAC and push the indoor CO_2 concentration toward 800 ppm, which is considered as the goal state. The architecture of the system is illustrated in Fig.1.

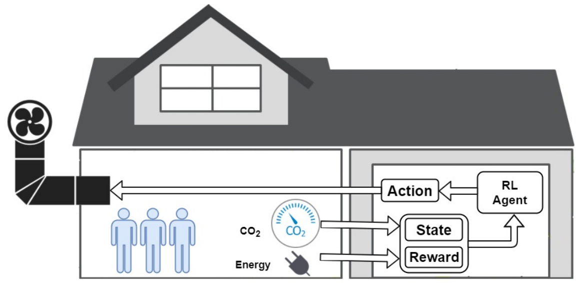

Fig. 1: Architecture of the ventilation system.

According to [12] for spaces where occupants vary with time and a mechanically ventilated system is used, the mass balance of CO_2 concentration can be expressed by equation (2). In this equation, V is space volume in m^3, $C(t)$ is the indoor CO_2 concentration at time t in ppm, $C_o(t)$ is outdoor CO_2 concentration at time t in ppm, $G(t)$ is CO_2 generation rate in the space at time t in L/s. Finally, Q is volumetric airflow rate into (and out of) the space in m^3/s.

$$V\frac{dc}{dt} = Q(C_o(t) - C(t)) + G(t) \quad (2)$$

Our approach provides a good indoor air quality by controlling the mechanical ventilation which is the result of the RL controller. According to (2), the indoor air quality is proportional to the occupants' number, ventilation rate and outdoor CO_2 concentration. The indoor CO_2 concentration is discretized from CO_2,min to $CO_{2,max}$ with integer values. Therefore there are $N = (CO_{2,max} - CO_{2,min} + 1)$ possible integer value which are named as the states of RL. The set of the possible states are shown as $\delta = \{\delta_1, \delta_2, \ldots, \delta_i, \ldots, \delta_{N-1}, \delta_N\}$ where $\delta_1 = CO_{2,min}$, $\delta_N = CO_{2,max}$ and $\delta_i = 1 + \delta_{i-1}$. It is assumed ventilation rate of the HVAC performs in M different mode, with M different energy consumption, where each one is known as an action of the RL. The set of the possible actions are shown as $\zeta = \{\zeta_1, \zeta_2, \ldots, \zeta_M\}$.

During the learning phase, the agent considers exploration as the way to learn user comfort and energy efficiency. According to [13], since CO_2 generation rates for the building occupants is varying slowly, therefore the CO_2 concentration is measured every 5 minutes. For any possible indoor CO_2 concentration, by applying the (3), embedded in the system, a comfort reward (r_{com}) is calculated. The aim of the comfort reward is to force the HVAC to meet the setpoint which is shown as $CO_{2,desire}$. At each timestamp t, if the agent takes an undesirable action at a certain state which takes the state far away from the $CO_{2,desire}$ stage, it gets a lower comfort reward. While if the agent performs the desired action which takes the state closer to the $CO_{2,desire}$ stage, a higher comfort reward associated with the state-action pair will be given. Consequently, in the future the agent tends to choose actions which gives a higher reward. The result of (3) is a comfort reward value, in the range of $[0, R]$, for an action-state pair of a and s at time t. In (3) $CO_{2,s'}$ is carbon dioxide concentration at state $s^{'}$ at time $t-1$ where $s \in \delta$, $s^{'} \in \delta$ and $a \in \zeta$. In this equation, W affects the intensity of comfort reward variation between two consecutive states $\delta_i \in \delta$ and $\delta_{i+1} \in \delta$.

$$r_{com}(s,a) = R \times \exp\{-\frac{(CO_{2,s'} - CO_{2,desire})^2}{W}\} \quad (3)$$

Another objective of our controller is minimizing the energy consumption of the HVAC system. An energy reward (r_e) based on the power saving due to going from one state to another state is calculated in every timestamp. Since different HVAC systems have different power consumption for the same action, consequently their power saving for the same action at the same state is different. Therefore, a normalized version of energy reward, (4), is applied for calculating the total reward (5). Results of energy reward are in the range of $[-1, 1]$, where $P(s)$, $P(s')$ and $P(\delta_n)$ is the power consumption at state $s \in \delta$, $s^{'} \in \delta$ and any state from 1 to N respectively.

$$r_e(s,a) = \frac{P(s^{'}) - P(s)}{\max_n P(\delta_n)} \tag{4}$$

Since the value of $r_{com}(s,a)$ and $r_e(s,a)$ have different scales, (5) is used to convert $r_e(s,a)$ range to the $r_{com}(s,a)$ range. The Scaled version of energy reward is shown as r_{en}.

$$\begin{aligned} r_{en}(s,a) = \frac{r_e(s,a) - \min(r_e)}{\max(r_e) - \min(r_e)} \times [\max(r_{com}) - min(r_{com})] \\ + \min(r_{com}) \end{aligned} \tag{5}$$

To make a decision, the agent calculates the total reward $r_{total}(s,a)$ which is a linear combination of comfort and energy rewards as follows, where, β is the Energy-Comfort balance factor, $0 < \beta < 1$.

$$r_{total}(s,a) = (1-\beta) \times r_{com}(s,a) + \beta \times r_{en}(s,a) \tag{6}$$

Since the power consumption of the HVAC at each ventilation rate is known, there is a priori knowledge about energy reward of (6) for each state-action pair. However, the comfort reward is started with an initial value and updated based on exploration, by applying (3).

IV. Case study

In order to analyze the proposed controller, we simulated a zone of $15m \times 10m \times 3m$, which it has the maximum capacity of 10 individuals. Minimum and maximum acceptable CO_2 concentration for the indoor are 750 and 1000 ppm respectively, while the desire CO_2 concentration is 800 ppm. Additionally, outdoor has a CO_2 concentration of 600 ppm. A 50-watt HVAC device, which operates with 0%, 4%, 8%, 12%, 16%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 100% of its maximum ventilation rate, $0.6m^3/s$, is assumed. To investigate the effect of the learning rate, (α), and the discount factor, (γ), of (1) on decisions we considered 0.3, 0.5, 0.8 values for α and γ. In (3) maximum comfort reward value, R, is set 40 and, W is selected as 30,000. Fig.2 demonstrate the comfort and energy reward variations of this scenario for acceptable CO_2 concentration range. To investigate the effect of comfort and energy rewards on the decisions, the "Energy-Comfort balance factor" (β) with values of 0, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.30, 0.35, 0.4, 0.45, 0.5, 0.8, 1 are applied in (6) and the total reward for different cases are calculated.

Fig.3 shows the convergence of Q-learning algorithm for different values of learning rate (α) and discount factor (γ). In this paper to conserve space, we just selected α and γ as 0.5 and 0.8 respectively, among the different combination of α and γ. For observing the effect of Energy-Comfort balance factor, (β), on total consumed energy, results for β equal to 0.4 and 0 are demonstrated. A duration of 24 hours is considered while the number of occupants changes during the simulation as shown in Fig.4.

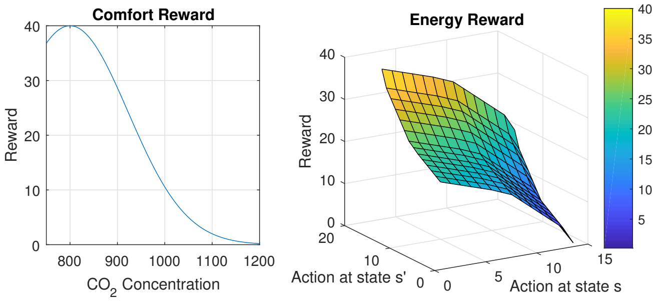

Fig. 2: Comfort reward and Energy reward.

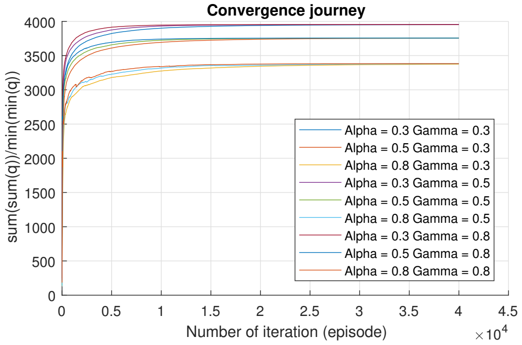

Fig. 3: Q-learning Convergence journey of case study for different value of and γ.

We compared our proposed HVAC Energy-Comfort control system based on RL with On-Off and setpoint (SP) control system. In On-Off control system, if any occupant is sensed in the operating zone of HVAC, it activates the ventilator with maximum power. The setpoint control system keeps the CO_2 at the desired setpoint by measuring the indoor CO_2 concentration without considering the existence of the occupant in the sensing zone. In this study the desired setpoint is selected as 800 ppm and ventilator operates with 0%, 4%, 8%, 12%, 16%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 100% of its maximum ventilation rate. The results related to CO_2 concentration and HVAC energy consumption are shown in Fig.4 and tableI.

By comparing the results of the investigated HVAC control systems, it is clear that the On-Off control method provides the fresh air during the simulation time, but consumes approximately five times more energy than the other methods. It is obvious that on-off control method activates the ventilator just when there is any occupant in the region. For the cases when the CO_2 concentration is over the acceptable or close to the upper band of the acceptable region when there is no individual in the region, the ventilator will not ventilate the indoor air. Therefore, in the next occupant hour, the quality of the indoor air will not start with good quality. The Setpoint control method with different ventilation rates provides an acceptable air quality during the simulation time in general. However, for the cases when the ventilator operates with more than 25% of its maximum rate, this method consumes more energy than our proposed method. In Setpoint control

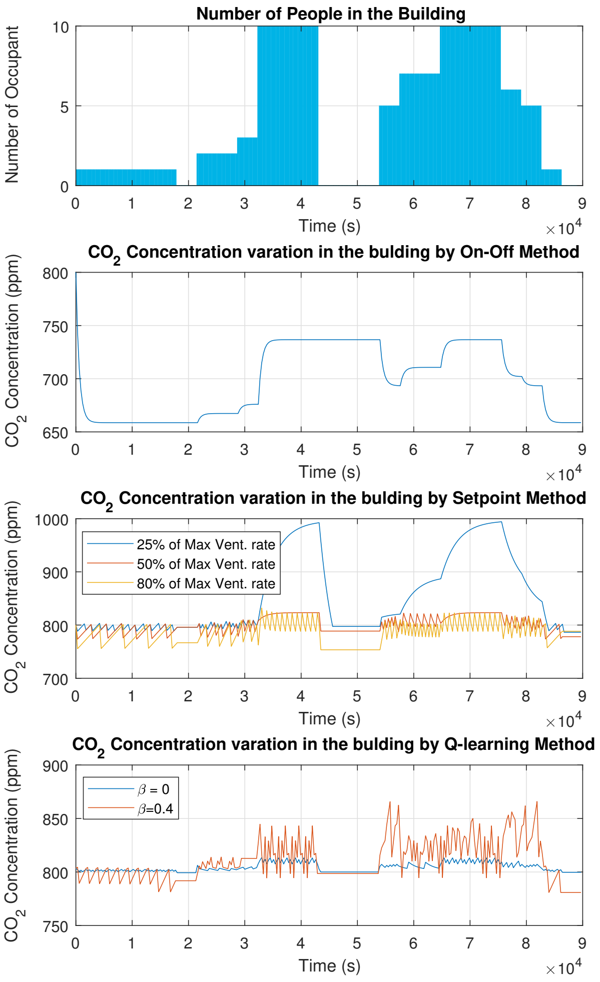

Fig. 4: CO_2 concentration during the simulation.

TABLE I: Comparison of power consumption and indoor CO_2 concentration for HVAC different control systems

	power consumption (kwatt)	Ave. indoor CO_2 during simulation (ppm)	Ave. indoor CO_2 during occupancy hours (ppm)
Q-learning method, $\alpha = 0.5, \gamma = 0.8, \beta = 0$	2.638	804.31	805.43
Q-learning method, $\alpha = 0.5, \gamma = 0.8, \beta = 0.4$	2.3125	810.02	814.11
On-Off method, 100% of max ventilator rate	12	697.98	696.10
Setpoint method, 25% of max ventilator rate	2.2125	852.58	863.72
Setpoint method, 50% of max ventilator rate	3.250	800.18	803.08
Setpoint method, 80% of max ventilator rate	3.360	784.91	790.15

method, if ventilator operates with less than or equal 25% of its maximum rate, it could not provide fresh air during the crowded occupant hours. Our proposed Q-learning method by considering the comfort and energy demands, consumes less energy than the other methods while providing the fresh air during the simulation. For $\beta = 0$ case, the HVAC consider the comfort reward solely. For $\beta = 0.4$ case, the comfort and energy reward are considered to calculate the total reward, while user comfort has a little more priority over than energy reward. Results of Fig.4 shows small variation of CO_2 concentration around 800 (ppm) for $\beta = 0$ case, while it consume more energy than $\beta = 0.4$ case. Since the proposed control system depends on CO_2 concentration and number of occupants, it can dynamically change its ventilator rates to satisfy the energy and comfort demands simultaneously.

V. Conclusion

This study focuses on applying Q-learning method, which intelligently considers the user comfort and energy efficiency and operates the HVAC system to keep the ventilation, i.e., the indoor CO_2 concentration, at the desired level. On-Off method and Setpoint methods with the possible ventilation rate of an HVAC system are simulated besides the proposed Q-learning control method under different occupant number in a building for the duration of 24 hours. The results proved the capability of the Q-learning control method to operate the HVAC energy efficiently and satisfied the occupants comfort simultaneously.

Acknowledgment

This work is part of the TEYDEB project, funded TUBITAK grant 7151322.

References

[1] Philipp Beiter, Michael Elchinger and Tian Tian, “2016 Renewable Energy Data Book,” Energy Efficiency Renewable Energy, U.S. Department of Energy, December 2017.

[2] Verhelst, C. “Model predictive control of ground coupled heat pump systems for office buildings”, PhD, Dep. Mechanical engineering, KU Leuven, Leuven, Belgium, 2012.

[3] Y. Ma, F. Borrelli, B. Hencey, B. Coffey, S. Bengea and P. Haves, "Model Predictive Control for the Operation of Building Cooling Systems," in IEEE Transactions on Control Systems Technology, vol. 20, no. 3, pp. 796-803, May 2012.

[4] K. H. Khan, C. Ryan and E. Abebe, "Day Ahead Scheduling to Optimize Industrial HVAC Energy Cost Based ON Peak/OFF-Peak Tariff and Weather Forecasting," in IEEE Access, vol. 5, pp. 21684-21693, 2017.

[5] L. Yu, T. Jiang and Y. Zou, "Online Energy Management for a Sustainable Smart Home with an HVAC Load and Random Occupancy," in IEEE Transactions on Smart Grid, vol. PP, no. 99, pp. 1-1, 2017.

[6] Reinforcement Learning : An Introduction (Richard S. Sutton, Andrew G. Barto), MIT Press, 1998.

[7] T. Hansen, E. Chong, S. Suryanarayanan, A. Maciejewski and H. Siegel, "A Partially Observable Markov Decision Process Approach to Residential Home Energy Management," in IEEE Transactions on Smart Grid, vol. PP, no. 99, pp. 1-1, June 2016.

[8] D. Zhang, S. Li, M. Sun and Z. O’Neill, "An Optimal and Learning-Based Demand Response and Home Energy Management System," in IEEE Transactions on Smart Grid, vol. 7, no. 4, pp. 1790-1801, July 2016.

[9] C. J. C. H. Watkins and P. Dayan, "Technical note: Q-learning," Mach. Learn., vol. 8, no. 3–4, pp. 279–292, May 1992.

[10] G.P. Henze and J. Schoenmann, “Evaluation of reinforcement learning control for thermal energy storage systems,” HVACR Research, Vol. 9, pp. 259-275, 2003.

[11] J. Vazquez-Canteli,J. Kampf,Z. Nagy, "Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted Q-iteration" in Energy Procedia, vol. 122, pp. 415-420, 2017.

[12] Xiaoshu Lu, Tao Lu and Martti Viljanen (2011), “Estimation of Space Air Change Rates and CO_2 Generation Rates for Mechanically-Ventilated Buildings, Advances in Computer Science and Engineering,” Dr. Matthias Schmidt (Ed.), ISBN: 978-953-307-173-2, InTech.

[13] Persily, A. and de Jonge, L, “ Carbon dioxide generation rates for building occupants,” in Indoor Air, vol. 27, no. 5, p-p. 868-879, 2017.