Thermal Comfort Control Based on MEC Algorithm for HVAC Systems

Dong Li, Dongbin Zhao, Yuanheng Zhu, Zhongpu Xia
The State Key Laboratory of Management and Control for Complex Systems. Institute of Automation,
Chinese Academy of Sciences. Beijing 100190, China
dongleecsu@gmail.com, dongbin.zhao@ia.ac.cn, zyh7716155@163.com, zhongpu.xia@gmail.com

***Abstract*—This paper combines an efficient reinforcement learning algorithm named Multisamples in Each Cell (MEC) with a building thermal comfort control problem. It implements the efficient exploration rule and makes high use of observed samples. A grid is utilized to partition the continuous state into cells that are used to store samples. A near-upper *Q* function is obtained based on the samples in each cell. The value iteration technique is designed to derive the near optimal control policy. The algorithm can efficiently balance exploration and exploitation. The entire implementation process needs no model of systems. The thermal comfort criterion, predicted mean vote, is introduced to evaluate zone thermal comfort status. A two story, multi-zone small office building equipped with a variable air volume direct expansion cooling system is built in EnergyPlus to establish an EnergyPlus-MATLAB co-simulation platform. A MEC thermal comfort control simulation is implemented to validate the high performance property compared with *Q*-learning.**

Keywords—MEC, Q-learning, thermal comfort control, EnergyPlus

I. Introduction

Energy shortage is one of the most severe problems around the world. In the USA, research shows that about 41% of the total energy consumption is from buildings [1]. The energy used in office sector is the largest in building sectors which include residential sector, office sector, and retail sector [2]. And the heating, ventilation, and air conditioning (HVAC) parts occupy nearly 40% of the office building's operating energy. The ultimate goal of using HVAC system is to improve the comfort sensation of the room. However, the traditional HVAC systems adopt simple temperature/humidity controllers, but neglect the thermal comfort control objective. Hence the improvement of the control strategy for HVAC, which has drawn much attention, is of great significance to our society. The green buildings, which are applied with efficient control scheme, can not only provide a better thermal environment for occupants, but also reduce the energy consumption and financial cost.

To implement the thermal comfort control, two basic elements--comfort criterion, and the model of HVAC systems are needed. Thermal comfort is the condition of mind that expresses satisfaction with the thermal environment, and is assessed by subjective evaluation [3]. Maintaining this standard of thermal comfort for occupants of buildings is the ultimate goal of using air conditioning. Therefore, the thermal comfort control is introduced as a key factor in this paper. In the past decades, several concepts of thermal comfort index have been proposed, such as predicted mean vote (PMV) and predicted percentage of dissatisfied (PPD). PMV, which is proposed by Fanger [4], evaluates zone comfort condition on a standard scale for a large group of persons. It is well accepted by the all world and adopted by ISO 7730 standard [5].

Many studies have been made based on the building HVAC system model. In the past few decades, there are many researches on building thermal control by using optimal approaches. Yahiaoui et al. [6] utilize linear programming to maintain indoor temperature in a comfortable level. [7] reviews supervisory control methods and optimization techniques used in the HVAC systems. All these studies above show a good thermal comfort performance of HVAC system by applying advanced control methods. However, among the HVAC system studies, a model of the HVAC system, which is used to build control scheme, is required. Developing a HVAC model is a tough task, which needs skillful mathematical knowledge and numerous sensor data of HVAC component and indoor, outdoor environment. This will definitely increase study difficulty and financial cost. So a model free control method can decrease the complexity of system, while saving financial cost.

Reinforcement learning [8] is a model free control method. It establishes a reward-punish mechanism by iterating the control strategy to realize the mapping from state to action. The mechanism is represented by numeric signal called reward. The agent is not told which actions to take, as in most forms of machine learning, but instead must discover which actions yield the highest reward by trial and error. [9, 10] present the active and passive building thermal storage inventory control by using reinforcement learning (RL), and demonstrate the good performance of the system.

Thanks to the development of building energy modeling tools, the HVAC simulation procedure becomes more convenient and accurate. For instance, EnergyPlus [11] is a validated and constantly updated energy simulation software. The simulation can be driven by the building parameters specified by the user. An update can be implemented

This work is supported by National Natural Science Foundation of China (NSFC) under Grants No. 61273136, and No. 61473176.

978-1-4799-1959-8/15/$31.00 ©2015 IEEE

throughout the building life-cycle when the details of the building changed. However, because of the lacking of the capability to interoperate the data with other simulation software like MATLAB, a toolbox named MLE+ [12] is needed to realize co-simulation between EnergyPlus and MATLAB. A building model is described in EnergyPlus, while control methods are specified in MATLAB. As a middle-ware, MLE+ can link two programs and exchange data between them.

In this paper, an EnergyPlus-MATLAB co-simulation platform is established. The contribution presents an efficient RL approach to solve the thermal comfort control problem based on the co-simulation platform. PMV criterion is applied to evaluate the comfort status of the room.

The structure of this paper is shown as follows. Section II introduces the thermal comfort criterion and presents the problem of thermal comfort control. The system framework and building modeling are described in Section III. RL background and MEC algorithm are introduced in Section IV. A simulation example is given to validate its effectiveness in Section V. Section VI draws the conclusion.

II. Problem Statement

As the condition of mind that expresses satisfaction with the thermal environment, thermal comfort can be evaluated by predicted mean vote (PMV). According to Fanger's theory [4], PMV can be calculated by combining air temperature, mean radiant temperature, relative humidity, air speed, metabolic rate, and clothing insulation. PMV can be defined as:

$$\begin{aligned} PMV = {} & (0.3033e^{-0.114M}+0.028)\times\{(M-W)-3.05[5.733- \\ & 0.000699(M-W)-Pa]-0.42[(M-W)-58.15]- \\ & 0.0173M(5.867-Pa)-0.0014M(34-Ta)-3.96\times \\ & 10^{8}\times fcl[(T_{cl}+273)^4-(T_{mrt}+273)^4]-fcl\times hc(T_{cl}-Ta)\} \end{aligned} \quad (1)$$

where

$$\begin{aligned} T_{cl} = {} & 35.7-0.028(M-W)-0.155I_{cl}\{3.96\times10^{-8} \\ & \times fcl[(T_{cl}+273)^4-(T_{mrt}+273)^4]-fcl\times h_c(T_{cl}-T_a)\}, \end{aligned} \quad (2)$$

$$h_c=\begin{cases} 2.38(T_{cl}-T_a)^{0.25}, & \text{if } 2.38(T_{cl}-T_a)^{0.25}\geq 12.1\sqrt{V_{air}} \\ 12.1\sqrt{V_{air}}, & \text{otherwise} \end{cases} . \quad (3)$$

In the above equations, M is metabolic rate ($\mathrm{W/m^2}$); W is external work, which equals to zero for most activities ($\mathrm{W/m^2}$); P_a is partial vapor pressure (Pa); fcl is the ratio of clothed body surface area to nude body surface area; T_{cl} is the surface temperature of clothing; I_{cl} is the thermal resistance of clothing (clo); h_c is the convectional heat transfer coefficient ($\mathrm{W/m^2\cdot K}$); V_{air} is the air velocity ($\mathrm{m^2/s}$).

The level of comfort is often characterized by using the ASHRAE thermal sensation scale, which is described in Table I. The ideal value is 0, which represents the most comfortable status. Therefore, the closer the PMV value is to 0, the better thermal comfort one will get.

TABLE I. ASHRAE Thermal Sensation Scale

PMV Value	**Sensation**
3	Hot
2	Warm
1	Slightly warm
0	Neutral
-1	Slightly cool
-2	Cool
-3	Cold

With (1), we can calculate PMV value in every step by varying zone air temperature and air velocity. The other variables like metabolic rate and external work can be set to constants that represent the average level, since we aim to control an office building thermal condition during the work time. According to [13], the metabolic rate can be set to 117 W/m^2 for the typing activity. Therefore, the input variables are HVAC temperature setpoint, which is used to control zone air temperature, and air velocity, which is used to adjust zone air flow speed. The output is zone PMV value that represents the thermal comfort condition. Reinforcement learning approach, like *Q*-learning [14], can learn to derive the optimal control policy by interacting with environment.

III. MEC for Thermal Comfort Control

In the basic RL problem, an agent interacts with environment and receives rewards. Four basic elements are included, such as the state s, the action a, the reward r, and the value function *V(s)* or the state-action value function *Q(s, a)*. Define *S* as states set and *A* as actions set. The policy *π*, which specifies the mapping from states set to actions set, can be denoted as $\pi : S \to A$. The controller chooses an action *a* following the policy π, then the environment will transform to a new state s' with reward *r*. The goal is to maximize the cumulative reward in a long term by tuning its policy.

A value function is defined to evaluate the performance of control policy π by summing the discounted reward in infinite horizon. We assume the initial state $s_0 = s$ and follow a policy π . The value function can be defined as:

$$V^{\pi}(s)=\sum_{k=0}^{\infty}\gamma^k r(s_k,a_k)\mid s_0=s, a_k=\pi_k(s_k) \quad (4)$$

where $\gamma\in(0,1)$ is the discount rate. The optimal policy π^* is the policy that can maximize the value of $V^{\pi}(s)$, i.e., $\pi^*=\arg\max_{\pi} V^{\pi}$. Sometimes the state-action value function is preferred, which is defined as $Q^{\pi}(s,a) \triangleq r(s,a) + \gamma V^{\pi}(s')$. Q function can also be described by applying the Bellman rule

$$Q(s,a) = Q(s,a) + \alpha(r(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)) \quad (5)$$

where $\alpha \in [0,1]$ is the learning rate. The optimal policy can be obtained by

$$\pi^{*} = \arg\max_{a \in A} Q^{\pi}(s,a). \quad (6)$$

In the thermal comfort problem, the state s which is rounded to the range of $\{-1.0, -0.9, \ldots, 1.3\}$ is the PMV value in every time step. Note that the state rang is large enough to cover the ordinary office PMV variant range because of the usage of air conditioning. The actions contain two dimensions, i.e., the temperature setpoint action $a^{(1)}$ and air velocity action $a^{(2)}$. The temperature setpoint (T_{sp}) is the discrete number in $\{21, 22, \ldots, 30\}$, while air velocity (V_{air}) is in $\{0, 0.05, ..., 0.5\}$. Therefore, the size of Q-table is $24 \times 10 \times 11$ in the Q-learning approach. In the current state s, the agent choose action $a = [a^{(1)}, a^{(2)}]$ with respect to policy π. Then, the state s will transfer to the next state s', and the agent will receive a reward r. According to [15], the optimal control policy, i.e., $\pi_i = \pi^{*}$, can be achieved if and only if $\pi_i(s) = \pi_{i+1}(s), \forall s \in S$. The value iteration [8] procedure can be implemented to obtain the optimal policy. However, in the traditional Q-learning approach, the training process is long due to the large state-action space, which means the learning efficiency is low. It is essential to utilize an efficient learning algorithm to overcome this drawback.

In order to explore efficiently for the finite Markov decision processes (MDPs), a new online RL, probably approximately correct (PAC), is developed in [16, 17]. PAC is an algorithm that the agent can learn a near optimal policy within a polynomial time or error bound. Among these algorithms, Zhao and Zhu [18] propose an approach named multi-samples in each cell (MEC) to approximate the near optimal state-action value function in the continuous MDPs. Compared with other RL algorithms, MEC is ensured to output a near-optimal policy, and the running time is finite and bounded, and it can make efficiently use of information of the system. Besides, the system dynamic is not needed by MEC. Hence as an efficient RL approach, MEC is suitable to the thermal comfort control problem.

The main principle of MEC is collecting observed samples into a data set selectively. The algorithm improves its performance based on these samples. The MEC algorithm includes three parts--data set construction, near-upper Q iteration, and escape event.

A. Data Set Construction

The continuous state space is divided into several small cells $C_i (1 \le i \le N_{grid})$ that contain some state action pairs and do not overlap with each other. The total number of cells in the grid is N_{grid}. We use $\Omega(C_i)$ to represent the state space in cell C_i. We can define the data set at time t as $D_t = \{s_k, a_k, r(s_k, a_k), s_{k+1}\}_{0 \le k \le t-1}$. The current time is t, and k is the previous time step. The observed state sample in time k is s_k, and $r(s_k, a_k)$ is the reward taking action a_k in state s_k. Therefore, any state action pair $(s_k, a_k) \in D_t$ is in a cell C_i. We apply $D_t(C_i, a)$ to describe the samples in D_t at a that belong to C_i. If there are no samples in data set D_t, we can denote this as $D_t(C_i, a) = \varnothing$.

B. Near-Upper Q Iteration

Given a function $g: S \times A \to \mathbb{R}$ and for any $s \in \Omega(C_i)$. The near-upper Q iteration (NUQI) operator $\overline{T}$ can be defined as following:

$$\overline{T} = \begin{cases} \dfrac{1}{K} \displaystyle\sum_{k: s_k \in D_t(C_i,a)} [r(s_k, a_k) + \gamma \max_{a'} g(s'_k, a')], & \\ & \text{if } D_t(C_i, a) \neq \varnothing \\ V_{\max}, & \text{otherwise} \end{cases} . \quad (7)$$

K is the total number of samples in data set D_t. The NUQI operator means that we set the upper bound of the value function $V_{\max}$ if corresponding C_i has no samples. According to [18], this technique can definitely encourage the sufficient exploration. For the thermal comfort control problem, the state s representing thermal comfort criterion PMV value, may gradually drift to the near state, though we take the same action a in the same state s. Therefore, the average is utilized to update the state-action value function, which can reduce the influence of drifting. Note that the original right side of NUQI operator in [18] is $\min_{\substack{k=1 \\ s_k \in D_t(C_i,a)}} [r(s_k, a_k) + \gamma \max_{a'} g(s'_k, a')]$, here we modify it to the average operation in (7) to eliminate the influence of drifting.

NUQI operator $\overline{T}$ has a fixed solution because it is a contraction. We assume $\overline{Q}_t$ is the fixed solution to (5) with respect to D_t. Value iteration method can be implemented to solve $\overline{Q}_t$. Note that for arbitrary state in cell C_i, they share the same samples set when calculating $\overline{Q}_t$ by (5). Hence there is no need to compute $\overline{Q}_t$ for all the states. Value iteration stops when the difference between two consecutive state value function is less than a positive number (e.g., 0.05 in this paper). Then, the greedy policy can be obtained by

$$\pi_t = \arg\max_{a\in A} \overline{Q}_t(s,a). \tag{8}$$

C. Escape Event

Once the greedy policy is obtained, one can implement the policy at current state s and transfer to the next state s'. However, whether the new state can provide new information about the system or not, one can refer to the following notion *known*. Assuming $s \in \Omega(C_i)$, the state action pair (s,a) is called *known* if and only if $D_t(C_i,a) \neq \varnothing$, and there exists a sample $(\hat{s},a,\hat{r},\hat{s}') \in D_t(C_i,a)$ such that $\hat{s}'$ and s' are in the same cell, where $\hat{s}'$ and s' are the next state when taking a in state $\hat{s}$ and s, respectively. Otherwise, the pair (s,a) is called *unknown*. If the sample at time t is known, then it can't bring new information and just is ignored. Otherwise, it is added to the data set D_t and (5) is solved to obtain the new solution $\overline{Q}_{t+1}$, which can be used to derive the new greedy policy π_{t+1} with respect to (6).

The escape event can be defined as: starting a trial from s and following policy π, an unknown pair (s_τ, a_τ) is encountered within $T_{\varepsilon/3}$ steps, where $t \le \tau \le t + T_{\varepsilon/3} - 1$. $T_{\varepsilon/3}$ is the $\varepsilon/3$-horizon time, where ε is the error bound. If the escape event occurs, it means that unknown pairs are encountered in the next $T_{\varepsilon/3}$ steps, so the agent is still learning. Therefore, the MEC algorithm can be obtained in Table II.

TABLE II. MEC ALGORITHM

```
Algorithm1: MEC Algorithm for Comfort Control
Require: $V_{\max}$ and a grid over state space $\{C_i\}$
Initialize: $D_\bullet \leftarrow \varnothing, \overline{Q}_\bullet \leftarrow V_{\max}, \pi_\bullet(s) = \arg\max_a \overline{Q}_\bullet(s,a)$
For $t = 0,1,2...$ do
  Observe $(s_k, a_k, r(s_k,a_k), s_{k+1})$
  If $(s_k,a_k)$ is unknown in $D_t$, then
    $(s_k, a_k, r(s_k,a_k), s_{k+1})$ is added to $D_t$
    Update $\overline{Q}_t$ according to NUQI
    Compute $\pi_t$ according to greedy rule
  End If
  Execute $\pi_t$ on agent
End For no escape event happens
```

IV. FRAMEWORK AND BUILDING MODELING

A. System Framework

Fig. 1 shows the framework of the thermal comfort control system. The control algorithm is implemented in MATLAB, which is connected with EnergyPlus by MLE+ tool box. The building features that are specified in EnergyPlus input file is shown in Table III. We give two input variables- T_{sp}, V_{air} to EnergyPlus, and the PMV value will be computed by (1). Then, the PMV value can be transmitted to thermal comfort controller via MLE+. If the reward is poor by taking above actions, which means that $Q(s,a)$ will decrease, the agent will find a better action in the rest actions with respect to $Q(s,a)$ in the next time step. Eventually, the agent can learn the optimal policy π^* by interacting with environment according to [18]. The controller can follow the optimal policy π^* and deliver the optimal action to EnergyPlus via MLE+.

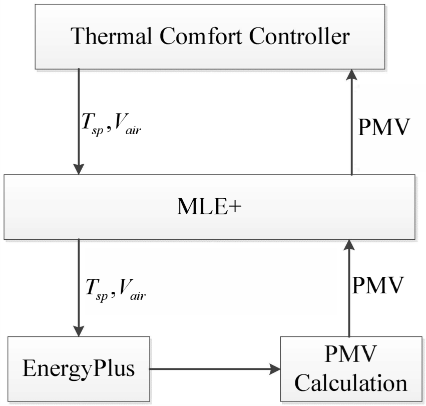

Fig. 1. Framework of simulation system.

TABLE III. MODEL FEATURES IN ENERGYPLUS

Floor area		618.7 m^2
Window to wall ratio		0.29
Internal loads	Occupant	13
	Lighting	12.1 kW
	Equipment	2750 W
Occupied hours		7:00 am ~ 18:00 pm
HVAC system		VAV Direct Expansion
Natural ventilation		None
Comfort variables		Metabolic rate: 117 W/m^2 Summer clothing resistance: 0.5 col Winter clothing resistance: 1.0 col

B. Building Modeling

The model built in this paper is a two story small office building shown in Fig. 2. Each floor includes two thermal zones: one north facing, and the other south facing. The details of model are listed in Table III.

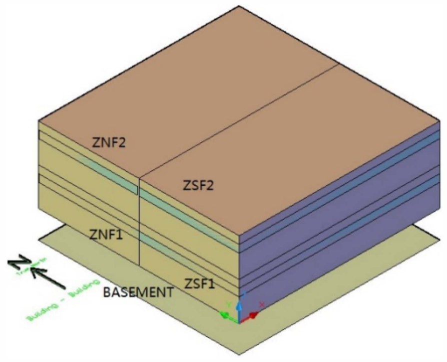

Fig. 2. 3-D Building model of small office.

V. Simulation

In this section, the MEC approach is applied to the building HVAC system described above. It will be compared with the Q-learning algorithm. A thermal comfort controller, which is controlled by MEC and Q-learning algorithm, respectively, is developed.

As described above, PMV can be utilized to judge the thermal comfort sensation. Therefore, the PMV value in every step is regarded as state s. The range of PMV value in this problem is around $[-1.0, 1.3]$, which is obtained from the prior knowledge of the HVAC system according to temperature setpoint and air velocity range described below. Note that the above PMV range is sufficiently large in the common place like office and home. Additionally, the PMV in $[-0.1, 0.1]$ is regarded as the same state, while other states are obtained by discretizing state space with step 0.1. The reason why the PMV in $[-0.1, 0.1]$ can be seemed as one state will be explained when we introduce reward definition. The total number of states is 24. In order to cover a large thermal comfort range, the temperature setpoint T_{sp} is set to the range of {21, 22…30}, and the second action-zone air velocity V_{air} is set to the range of {0, 0.05…0.5}. So the actions space is a matrix with the size of 10 by 11. According to the meaning of PMV value, zero is the ideal value that represents the most comfortable sensation. The closer the PMV value is to zero, the better thermal comfort status we get. Basing this rule, the reward can be defined as:

$$r = \begin{cases} 0, & \text{if } |PMV| < 0.1 \\ -200(|PMV| - 0.1)^2, & \text{otherwise} \end{cases}. \qquad (9)$$

One can tell from (9) that the control goal is to keep PMV value in [-0.1, 0.1] by changing the temperature setpoint. Any action leading the PMV value out of this range will be punished with different degree.

The two control algorithms are implemented in MATLAB, which is connected with EnergyPlus by MLE+ tool box. As the data exchange layer, MLE+ receives a two dimensional action vector and transmits it to EnergyPlus by using building controls virtual test bed (BCVTB). A new simulation round is implemented by utilizing the new action following policy π. The output of the new round will feedback to controller by MLE+ to update the Q function by (5) or (7). The HVAC system is on from 7:00 to 18:00 at June 1st, and the time step is 1 minute (EnergyPlus inner time line).

The parameters of two control approaches are as follows. In the MEC algorithm, $V_{\max}$ is set to 0. In Q-learning algorithm, the learning rate α is 0.1. The exploration probability ε is initialized by 0.25 and decreases by 0.005 per episode. The discount rate γ is set to 0.98. The initial state of each episode is chosen randomly from states space, while the Q-table continues using the last one. Each episode contains 660 steps.

Both approaches are trained offline. Fig. 3 illustrates the convergence process via the summation of discounted rewards, i.e., the value function. The solid blue line is the MEC training process, while the dashed red line refers to the Q-learning's. One can tell from Fig. 3 that MEC algorithm converges around the 33rd episode with a higher value function which equals to -123.7. The traditional Q-learning algorithm converges slower than MEC and yields a lower value function around -139.5.

We assume the system has been trained according to the above training process. Set the initial PMV as 0.75. Fig. 4 shows the thermal comfort performance of the two control approaches. One can tell that the MEC algorithm can control the PMV value to the range [-0.1, 0.1] in 4 minutes. However, it takes about 12 minutes to get into that range by using Q-learning algorithm. The green lines in the Fig. 4 are the upper and lower bound of the comfortable sensation.

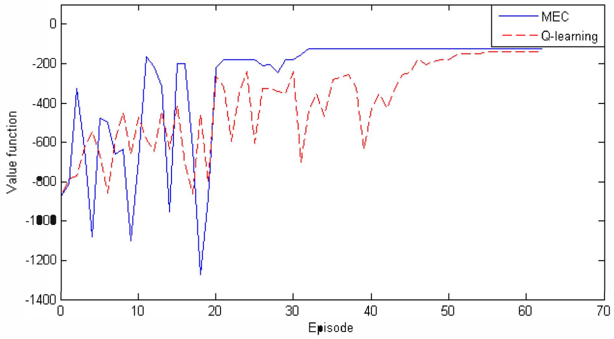

Fig. 3. The convergence process of thermal comfort control.

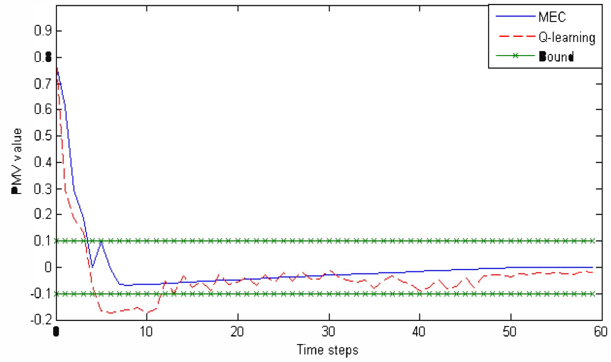

Fig. 4. The thermal comfort performance from state 0.75.

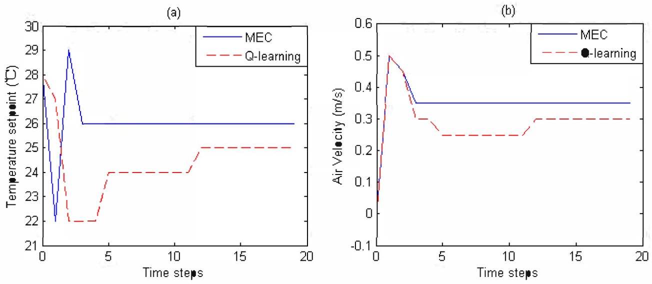

Fig. 5. The action trajectories from state 0.75.

The corresponding trajectories of temperature setpoint (a) and air velocity (b) are described in Fig. 5. The actions of two approaches are stable on different actions because of the different policy. As to the stable speed, MEC method is faster than *Q*-learning.

Note that change the initial status will not influence the good performance of MEC approach. Set the initial PMV as 0.4 and utilize the trained controller to handle the thermal comfort status. Fig. 6 illustrates the thermal comfort performance of the two control approaches. The MEC approach can also reach the goal faster than *Q*-learning. The corresponding action trajectories are shown in Fig. 7.

All these results indicate that the MEC method can obtain not only higher convergent speed, but also the better value function compared with *Q*-learning method.

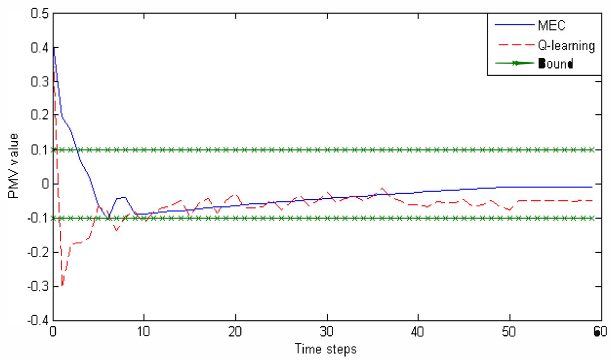

Fig. 6. The thermal comfort performance from state 0.4.

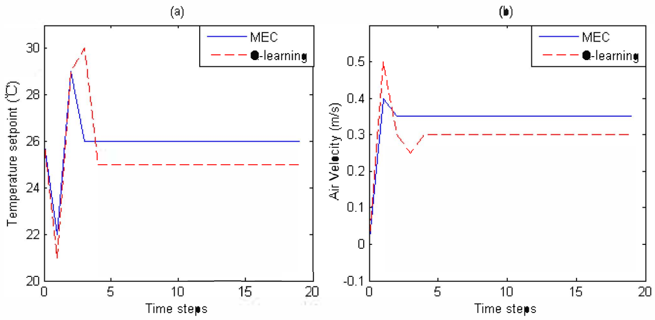

Fig. 7. The action trajectories from state 0.4.

VI. Conclusion

In this paper, MEC approach is presented for a building HVAC system without accessing the dynamic model. PMV criterion is introduced to analyze the variables affecting zone thermal comfort condition. Value iteration approach is implemented to obtain the optimal control policy. A small office building HVAC system is built in EnergyPlus. The MLE+ toolbox is utilized as a middleware to link EnergyPlus and MATLAB. MEC and *Q*-learning controller are trained in MATLAB with the data gained from EnergyPlus through MLE+.

A building HVAC system thermal comfort control simulation is implemented to test the effectivity and feasibility of the control method. This example illustrates that MEC approach can realize the optimal comfort control and keep the zone sensation into a good status. Additionally, the learning speed of MEC is higher than the *Q*-learning approach, and the thermal comfort performance is better than the *Q*-learning's. Hence the example demonstrates that the MEC approach is more effective for building thermal comfort control.

Future work will be on the extension to multi-objective control of the building energy system with respect to energy consumption analysis.

References

[1] EIA, "Annual energy review 2010," Annual report, The World Business Council For Sustainable Development, October 2011.

[2] Nguyen T A, Aiello M, "Energy intelligent buildings based on user activity: a survey," Energy and Buildings, 2013, 56: 244-257.

[3] "Thermal environmental conditions for human occupancy," ANSI/ASHRAE Standard 55-2013.

[4] P. O. Fanger, "Thermal comfort," Danish Technical Press, Copenhagen, 1970.

[5] "Moderate thermal environment-determination of the pmv and ppd indices and specification of the conditions for thermal comfort," International standard ISO 7730, 1994

[6] A. Yahiaoui, J. Hensen, L. Soethout, D. van Paassen, "Model based optimal control for intergrated building systems," 6th Int. Postgraduate Research Conf. in the Built and Human Environment, 2006: 322-332.

[7] S. Wang, Z.Ma, "Supervisory and optimal control of building HVAC systems: a review," HVAC&R Research, 2008, 14(1): 3-32.

[8] R. S. Sutton, A. G. Barto, "Reinforcement learning: an introduction," Cambridge, MA, USA: MIT press, 1998, no. 1.

[9] S. Liu, G. P. Henze, "Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: part 1. theoretical foundation," Energy and Buildings, 2006, 38(2): 142-147.

[10] S. Liu, G. P. Henze, "Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: part 2: results and analysis," Energy and Buildings, 2006, 38(2): 148-161.

[11] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann, et al, "EnergyPlus: creating a new-generation building energy simulation program," Energy and Buildings, 2001, 33(4): 319-331.

[12] W. Bernal, M. Behl, T. X. Nghiem, R. Mangharam, "MLE+: a tool for integrated design and deployment of energy efficient building controls," in Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, 2012: 123-130.

[13] "EnergyPlus input output reference," http://apps1.eere.energy.gov/buildings/energyplus/energyplus_documentation.cfm

[14] C. J. C. H. Watkins, P. Dayan, "Q-learning," Machine learning, 1992, 8(3-4): 279-292.

[15] D. B. Zhao, Z. P. Xia, D. Wang, "Model-free optimal control for affine nonlinear systems based on action dependent heuristic dynamic programming with convergency analysis," IEEE Trans. Automation and Science Engineering, DOI.10.1109/TASE.2014.2348991.

[16] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," Machine Learning, 2002, (49): 209-232.

[17] A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman, "PAC model-free reinforcement learning," in Proc. 23rd Int. Conf. Mach. Learn., 2006, pp. 881-888.

[18] D. B. Zhao, Y. H. Zhu, "MEC-a near-optiaml online reinforcement learning algorithm for continuous deterministic system," IEEE Trans. Neural Networks and Learning Systems, 2015, 26(2): 346-356.