Hindawi Publishing Corporation
International Journal of Photoenergy
Volume 2015, Article ID 496401, 12 pages
http://dx.doi.org/10.1155/2015/496401

Research Article

A Reinforcement Learning-Based Maximum Power Point Tracking Method for Photovoltaic Array

Roy Chaoming Hsu,[1] Cheng-Ting Liu,[2] Wen-Yen Chen,[2] Hung-I Hsieh,[1] and Hao-Li Wang[2]

[1]*Department of Electrical Engineering, National Chiayi University, Chiayi City 60004, Taiwan*
[2]*Department of Computer Science and Information Engineering, National Chiayi University, Chiayi City 60004, Taiwan*

Correspondence should be addressed to Hung-I Hsieh; hihsieh@mail.ncyu.edu.tw and Hao-Li Wang; haoli@mail.ncyu.edu.tw

Received 28 November 2014; Accepted 20 March 2015

Academic Editor: Marcelo C. Cavalcanti

A reinforcement learning-based maximum power point tracking (RLMPPT) method is proposed for photovoltaic (PV) array. By utilizing the developed system model of PV array and configuring the environment for the reinforcement learning, the proposed RLMPPT method is able to observe the environment state of the PV array in the learning process and to autonomously adjust the perturbation to the operating voltage of the PV array in obtaining the best MPP. Simulations of the proposed RLMPPT for a PV array are conducted. Experimental results demonstrate that, in comparison to an existing MPPT method, the RLMPPT not only achieves better efficiency factor for both simulated weather data and real weather data but also adapts to the environment much fast with very short learning time.

1. Introduction

The U.S. Energy Information Administration (EIA) estimates that the primary sources of energy consisted of petroleum, coal, and natural gas, amounting to over 85% share for fossil fuels in primary energy consumption in the modern world. Yet, recent years' over exploitation and consumption, and the expectation of depletion of fossil fuel, bring energy crisis to the modern world. Besides, awareness of environmental protection and sustainability in burning of fossil fuel and its products as the primary energy source are also arising. Many environment researchers and environmentalists advocated energy conservation and carbon dioxide (CO_2) reduction for the well-being of earth creatures and humans as well. As such, many alternatives for energy, such as energy generated from geothermal, solar, tidal, wind, and waste, are suggested. Among these, solar energy is the most used and promising alternative energy with a fast growing energy market share in the world's energy industry due to the following advantages.

(i) The sunlight and heat for the generation of solar energy is inexhaustible.

(ii) Sunlight is easy to access for its irradiance covering the most of the land.

(iii) There is no noise or pollution in the generation of solar energy.

(iv) Solar energy is considered as safe energy without burning any material.

Owing to the above advantages, many countries in the world started to establish energy policy and develop the related industries for solar energy since 70s.

Solar energy is normally generated by utilizing a photovoltaic electrical device, called solar cell or photovoltaic cell, in converting the energy of sunlight into electrical energy. Solar cells may be integrated to form modules or panels, and large photovoltaic arrays may also be formed from the panels. The performance of a photovoltaic (PV) array system depends on the solar cell and array design quality and on the operating conditions as well. The output voltage, current, and power of PV array vary as functions of solar irradiation level, temperature, and load current. Hence, in the design of PV arrays, the PV array output to the load/utility should not be adversely affected by the change in temperature and solar irradiation levels. On the other hand, improvement of the conversion efficiency of the PV array is an issue worth exploring. Generally speaking, there

are three means to improve the efficiency of photoelectric conversion: (1) increasing the photoelectric conversion efficiency of photovoltaic diode components, (2) increasing the frequency of direct light, and (3) improving the maximum power point tracking (MPPT) for the PV array. The first and second methods are to improve the hardware devices, yet the third one is to improve the conversion efficiency by utilizing the internal software embedded in the PV array system, which attracts many attentions. Hence, many MPPT methods have been proposed [1], like perturbation and observation method [2–4], the open-circuit voltage method [5], the swarm intelligence method [6], and so on.

In this paper, a reinforcement learning-based MPPT (RLMPPT) method is proposed to solve the MPPT problem for the PV array. In the RLMPPT, after observing the environmental conditions of the PV array, the learning agent of the RLMPPT determines the perturbation to the operating voltage of the PV array, that is, the action, and receives a reward by the rewarding function. By receiving rewards, the RLMPPT is encouraged to select (state, action) pairs with positive rewards. Hence, a series of actions with received positive rewards is generated iteratively such that a (state, action) pair selection strategy is gradually achieved in the so-called "learning" process. Once the agent of the RLMPPT learned the strategy, it is able to autonomously adjust the perturbation to the operating voltage of the PV array to obtain the maximum power for tracking the MPPT of the PV array. Research contributions of this study are summarized as follows:

(i) The proposed RLMPPT solves the MPPT problem of PV array with reinforcement leaning method, which is novel, to the best of our knowledge, to the area of MPPT of a PV system.

(ii) Reward function constructed from the early MPP knowledge of a PV array, experienced from past weather data, is employed in the learning process without predetermined parameters required by certain MPPT techniques.

(iii) Comprehensive experimental results exhibit the advantage of the RLMPPT in self-learning and self-adapting to varied weather conditions for tracking the maximum power point of the PV array.

The rest of the paper is organized as follows. In Section 2, we present the concept of MPPT for PV systems. Section 3 introduces the proposed RLMPPT for the PV array. The experimental configurations are described in Section 4. In Section 5, the results are illustrated with figures and tables. Finally, Section 6 concludes the paper.

2. Concepts of MPPT for PV Systems

2.1. Review of Operating Characteristics of Solar Cell. Solar cells are typically fabricated from semiconductor devices which produce DC electrical power when they are exposed to sunlight of adequate energy. When the cells are illuminated by solar photons, the incident photons can break the bonds of ground-state (valence-band, at a lower energy level) electrons, so that the valence electrons can then be pumped by those photons from the ground-state to the excited-state (conduction-band, at a higher energy level). Therefore, the free mobile electrons are driven to the external load, to generate the electrical power via a wire, and then are returned to the ground-state at a lower energy level. Basically, an ideal solar cell can be modeled by a current source in parallel with a diode; however, in practice, a real solar cell is more complicated and contains a shunt and series resistances R_{sh} and R_s. Figure 1(a) shows an equivalent circuit model of a solar cell, including the parasitic shunt and series elements, in which a typical characteristic of practical solar cell with neglecting the R_{sh} can be described by [7–10]

$$I_{pv} = I_{ph} - I_{pvo}\left\{\exp\left[\frac{q}{AkT}\left(V_{pv} + I_{pv}R_s\right)\right] - 1\right\}, \quad (1)$$

$$V_{pv} = \frac{q}{AkT}\ln\left(\frac{I_{ph} - I_{pv} + I_{pvo}}{I_{pvo}}\right) - I_{pv}R_s, \quad (2)$$

where I_{ph} is the light-generated current, I_{pvo} is the dark saturation current, I_{pv} is the PV electric current, V_{pv} is the PV voltage, R_s is the series resistance, A is the nonideality factor, k is the Boltzmann constant, T is the temperature, and q is the electron charge. The output power from PV cell can then be given by

$$P_{pv} = V_{pv}I_{pv} = I_{pv}\left\{\frac{q}{AkT}\ln\left(\frac{I_{ph} - I_{pv} + I_{pvo}}{I_{pvo}}\right) - I_{pv}R_s\right\}. \quad (3)$$

The above equations can be applied to simulate the characteristics of a PV array provided that the parameters in the equations are known to the user. Figure 1(b) illustrates the current-voltage (I-V) characteristic of the open-circuit voltage (V_{oc}), the short-circuit current (I_{sc}), and the power operation of a typical silicon solar cell.

As can be seen in the figure, the parasitic element R_{sh} has no effect on the current I_{sc}, but it decreases the voltage V_{oc}; in turn, the parasitic element R_s has no effect on the voltage V_{oc}, but it decreases the current I_{sc}. According to (1)–(3), a more accurate representation of a solar cell under different irradiances, that is, the current-to-voltage (I_{pv}-V_{pv}) and power-to-voltage (P_{pv}-V_{pv}) curves, can be described in the same way with different levels as shown in Figure 2. The maximum power point (MPP) in this manner occurs when the derivative of the power P_{pv} to voltage V_{pv} is zero, where

$$\frac{dP_{pv}}{dV_{pv}} = 0. \quad (4)$$

The resulting I-V and P-V curves presented in such way are shown in Figure 2.

2.2. Review of MPPT Methods. The well-known perturbation and observation (P&O) method for PV MPP tracking [2–4] has been extensively used in practical applications because the idea and implementation are simple. However, as reported by [11, 12], the P&O method is not able to track

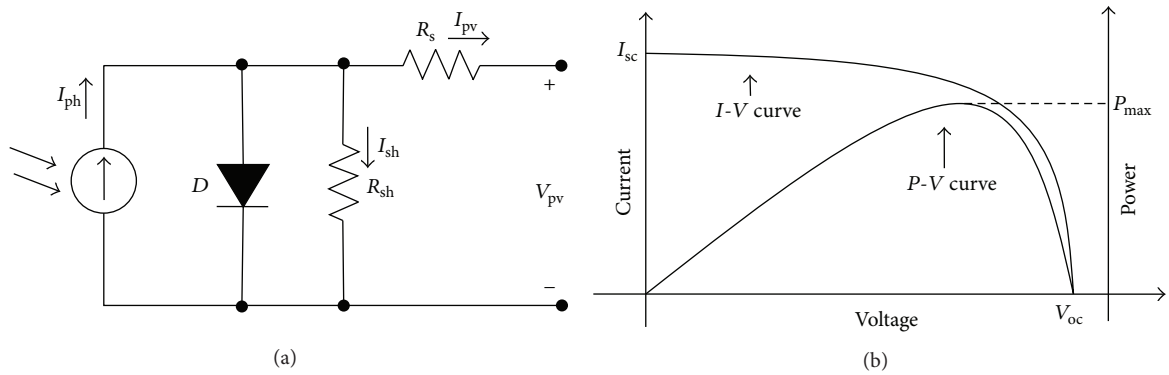

FIGURE 1: (a) Equivalent circuit model of a solar cell. (b) Solar cell I-V and power operation curve with the characteristic of V_{oc} and I_{sc}.

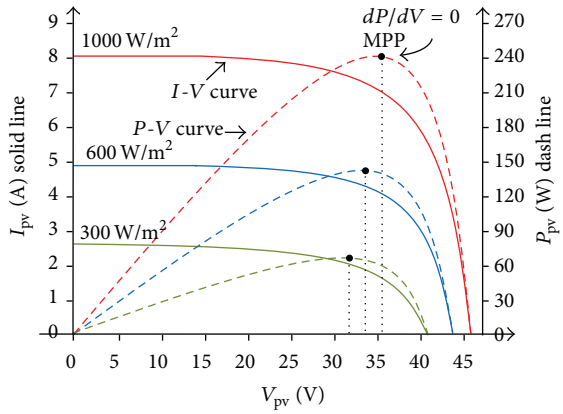

FIGURE 2: The I_{pv}-V_{pv} and P_{pv}-V_{pv} characteristics of different irradiance levels. The maximum power point (MPP) occurring when the derivative dP/dV is zero.

peak power conditions during periods of varying insolation. Basically, P&O method is a mechanism to move the operating point toward the maximum power point (MPP) by increasing or decreasing the PV array voltage in a tracking period. However, the P&O control always deviates from the MPP during the tracking, whose behavior results in oscillation around MPP in case of constant or slowly varying atmospheric conditions. Although this issue can be improved by further decreasing the perturbation step, the tracking response will become slower. Under rapidly changing atmospheric conditions for PV array, the P&O method may sometimes make the tracking point far from the MPP [13, 14].

2.3. Estimation of MPP by Using V_{oc} and I_{sc}. The open-circuit voltage V_{oc} and short-circuit current I_{sc} of the PV panel can be measured, respectively, when the terminal of the PV panel is open or short. In reality, both V_{oc} and I_{sc} are seriously dependent on the solar insolation. However, the maximum power point (MPP) is always located around the roll-off portion of the I-V characteristic curve in any insolation. Interestingly, there at MPP appears certain relation between the MPP set (I_{mpp}, V_{mpp}) and the set (I_{sc}, V_{oc}), which is worthy of studying. Further, the mentioned relation by empirical estimation seems always to hold and not to be subject to the insolation variation. It can be presumed, from commonly knowledge of PV array, that, in open-circuit mode, the relation of V_{mpp} and V_{oc} will be

$$V_{mpp} = k_1 V_{oc}, \tag{5}$$

and, in the short-circuit mode, the relation of I_{mpp} and I_{sc} will be

$$I_{mpp} = k_2 I_{sc}, \tag{6}$$

where k_1 and k_2 are constant factors between 0 and 1. From (5) and (6), we have the maximum power at MPP; that is,

$$P_{mpp} = V_{mpp} I_{mpp} = k_1 k_2 V_{oc} I_{sc}. \tag{7}$$

Even the P_{mpp} for learning is point estimation; it should be given by satisfying the MPP criteria in (4). For learning, the empirical result shows that the initial factor k_1 for V_{mpp} is around 0.8 and the k_2 for I_{mpp} is around 0.9.

3. The Proposed RLMPPT for PV Array

3.1. Reinforcement Learning (RL). RL [15–17] is a heuristic learning method that has been widely used in many fields of application. In the reinforcement learning, a learning agent learns to achieve the predefined goal mainly by constantly interacting with the environment and exploring the appropriate actions in the state the agent situates. The general model of reinforcement learning is shown in Figure 3, which includes the agent, environment, state, action, and reward.

The reinforcement learning is modeled by the Markov decision process (MDP), where a RL learner, referred to as an agent, consistently and autonomously interacts with the MDP environment by exercising its predefined behaviors. A MDP environment consists of a predefined set of states, a set of controllable actions, and a state transition model. In general, the first order MDP is considered in the RL, where the next state is only affected by the current state and action. For the cases where all the parameters of a state transition model are known, the optimal decision can be obtained by using dynamic programming. However, in some real world

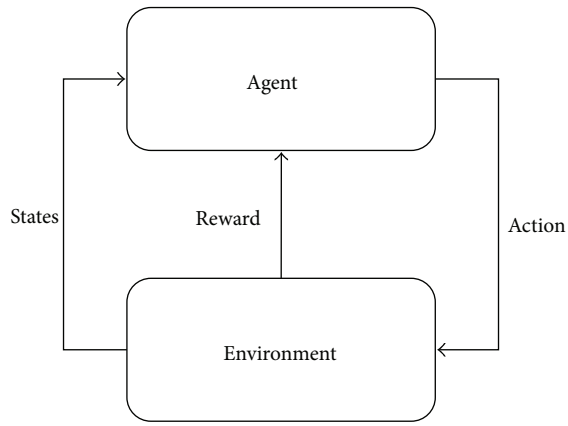

Figure 3: Model of the reinforcement learning.

cases the model parameters are absent and unknown to the users; hence, the agent of the RL explores the environment and obtains reward from the environment by try-and-error interaction. The agent then maintains the reward's running average value for a certain state-action pair. According to the reward value, the next action can be decided by some exploration-exploitation strategies, such as the ε-greedy or softmax [15, 16].

Q-learning is a useful and compact reinforcement learning method for handling and maintaining running average of reward [17]. Assume an action a is applied to the environment by the agent and the state goes to s' from s and receives a reward r'; the Q-learning update rule is then given by

$$Q(s,a) \longleftarrow Q(s,a) + \eta \Delta Q\left(s', a\right), \tag{8}$$

where η is the learning rate for weighting the update value to assure the convergence of the learning, the $Q(\cdot)$ is the reward function, and the delta-term, $\Delta Q(\cdot)$, is represented by

$$\Delta Q(s,a) = \left[r' + \gamma Q^*\left(s'\right)\right] - Q(s,a), \tag{9}$$

where r is the immediate reward; γ is the discount rate to adjust the weight of current optimal value, $Q^*(\cdot)$, whose value is computed by

$$Q^*\left(s'\right) = \max_{\forall b \in \mathbf{A}} Q\left(s', b\right). \tag{10}$$

In (10), $\mathbf{A}$ is the set of all candidate actions. The learning parameters of η and γ, in (7), and (8), respectively, are usually set with value ranges between 0 and 1. Once the RL agent successfully reaches the new state s', it will receive a reward r' and update the Q-value; then, the s' is substituted by the next state, that is, $s \leftarrow s'$, and the upcoming action is then determined according to the predefined exploration-exploitation strategy, such as the ε-greedy of this study. And the latest Q-value of the state is applied to the environment from one state to the other.

3.2. State, Action, and Reward Definition of the RLMPPT. In the RLMPPT, the agent receives the observable environmental signal pair of $(V_{\text{pv}}(i), P_{\text{pv}}(i))$, which will be subtracted from the previous signal pair in obtaining the $(\Delta V_{\text{pv}}(i), \Delta P_{\text{pv}}(i))$ pair. $\Delta V_{\text{pv}}(i)$ and $\Delta P_{\text{pv}}(i)$ each takes positive or negative signs to constitute a *state* vector, $\mathbf{S}$, with four states. The agent then adaptively decides and executes the desired perturbation $\Delta v(i)$, characterized as *action* to the V_{pv}. After the selected action is executed, a reward signal, $r(i)$, is calculated and granted to the agent; and, accordingly, the agent then evaluates the performance of the state-action interaction. By receiving the rewards, the agent is encouraged to select the action with the best reward. This leads to a series of actions with the best rewards being iteratively generated such that MPPT tracking with better performance is gradually achieved after the learning phase. The state, action, and reward of the RLMPPT for the PV array are sequentially defined in the following.

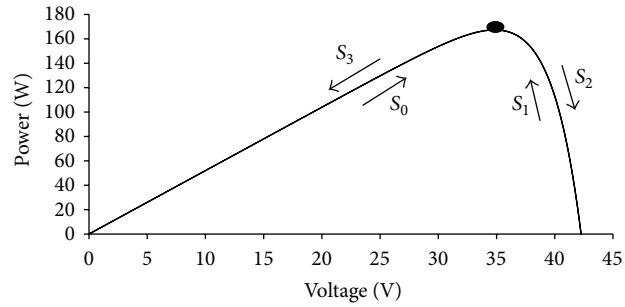

Figure 4: Four states of the RLMPPT.

(i) States. In RLMPPT, the state vector is denoted as

$$S = \left[S_0, S_1, S_2, S_3\right] \subseteq \mathbf{S}, \tag{11}$$

where $\mathbf{S}$ is the space of all possible environmental state vectors with elements transformed from the observable environment variables, $\Delta V_{\text{pv}}(i)$ and $\Delta P_{\text{pv}}(i)$, where S_0, S_1, S_2, and S_3, respectively, represent at any sensing time slot the state of going toward the MPP from the left, the state of going toward the MPP from the right, the state of leaving from the MPP to the left, and the state of leaving from the MPP to the right. The four states of the RLMPPT can be shown in Figure 4, where S_0 indicates that $\Delta V_{\text{pv}}(i)$ and $\Delta P_{\text{pv}}(i)$ have all the positive sign, S_1 indicates that $\Delta V_{\text{pv}}(i)$ is negative and the $\Delta P_{\text{pv}}(i)$ is positive sign, S_2 indicates that $\Delta V_{\text{pv}}(i)$ is positive and $\Delta P_{\text{pv}}(i)$ is negative sign, and finally S_3 indicates that $\Delta V_{\text{pv}}(i)$ and $\Delta P_{\text{pv}}(i)$ are all the negative sign.

(ii) Actions. The action of the RLMPPT agent is defined as the controllable variable of the desired perturbation $\Delta v(i)$ to the $V_{\text{pv}}(i)$, and the state of the agent's action, A_{per}, is denoted by

$$A_{\text{per}} \in \mathbf{A} = \left\{d_0, d_1, \ldots, d_N\right\}, \tag{12}$$

where $\mathbf{A}$ is a set of all the agent's controllable perturbations $\Delta v(i)$ for adding to the $V_{\text{pv}}(i)$ in obtaining the power from the PV array.

(iii) Rewards. In RLMPPT, rewards are incorporated to accomplish the goals of obtaining the MPP of the PV array.

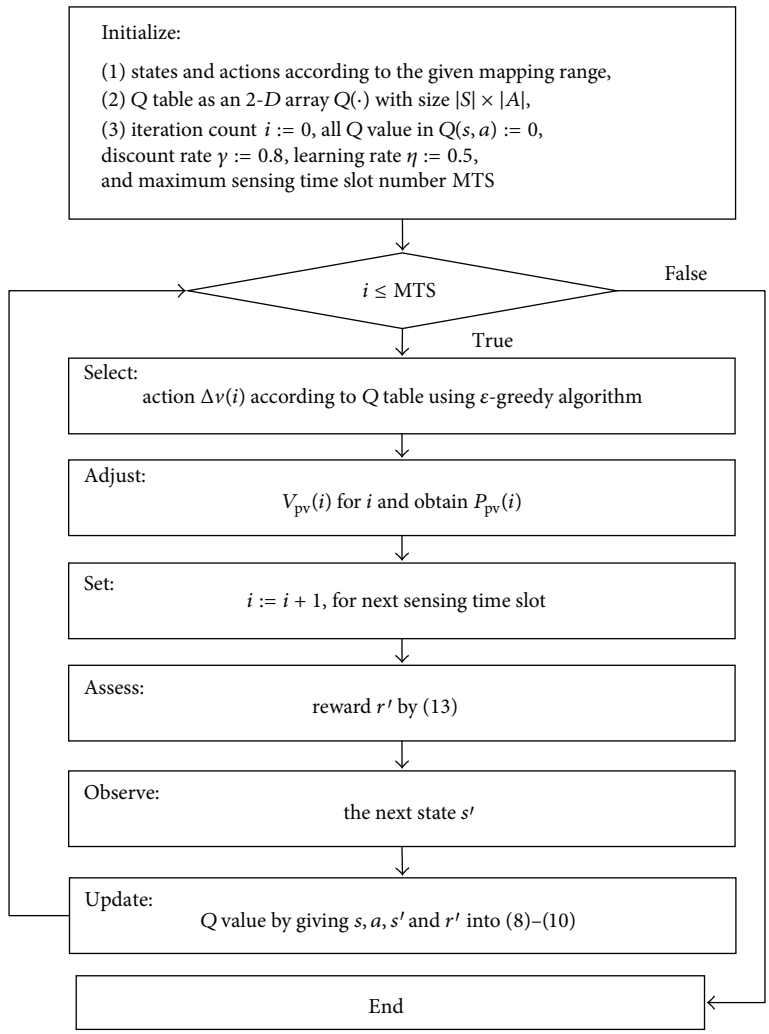

Figure 5: The flowchart of the proposed RLMPPT for the PV array.

Intuitively, the simplest but effective derivation of reward could be a *hit-or-miss* type of function, that is, once the observable signal pair $(V_{pv}(i), P_{pv}(i))$ hits the sweetest *spot*, that is, the true MPP of $(V_{mpp}(i), P_{mpp}(i))$, a positive reward is given to the RL agent; otherwise, zero reward is given to the RL agent. The *hit-or-miss* type of reward function could intuitively be defined as

$$r_{i+1} = \delta\left(\left(V_{pv}(i), P_{pv}(i)\right), h\_z(i)\right), \tag{13}$$

where $\delta(\cdot)$ represents the *Kronecker delta* function, and $h\_z(i)$ is the hitting spot defined for the i_{th} sensing time slot. In defining (13), a negative reward explicitly represents punishment for the agent's failure; that is, missing the hitting spot could achieve better learning results in comparison with a zero reward in the tracking of MPP of a PV array. In reality, the possibility that the observable signal pair $(V_{pv}(i), P_{pv}(i))$, leaded by the agent's action of perturbation $\Delta v(i)$ to the $V_{pv}(i)$, exactly hits the sweetest spot of MPP is very low in any sensing time slot i. Besides, it is also very difficult for the environment's judge to define a hitting spot for every sensing time slot. Hence, the hitting spot $h\_z(i)$ in (13) can be relaxed where a required *hitting zone* is defined on the previous environmental knowledge on a diurnal basis and where a positive/negative reward will be given if $(V_{pv}(i), P_{pv}(i))$ falls into/outside the predefined hitting zone in any time slot. Hence, the reward function can be formulated as

$$r_{i+1} = \begin{cases} c_1, & \left(V_{pv}(i), P_{pv}(i)\right) \in \text{Hitting Zone} \\ -c_2, & \text{otherwise,} \end{cases} \tag{14}$$

where c_1 and c_2 are positive values, denoting weighting factors in maximizing the difference between reward and punishment for better learning effect. In this study, the ε-greedy algorithm is used in selecting the RLMPPT agent's actions to avoid repeatedly selection of the same action. The flowchart of the proposed RLMPPT of this study is shown in Figure 5.

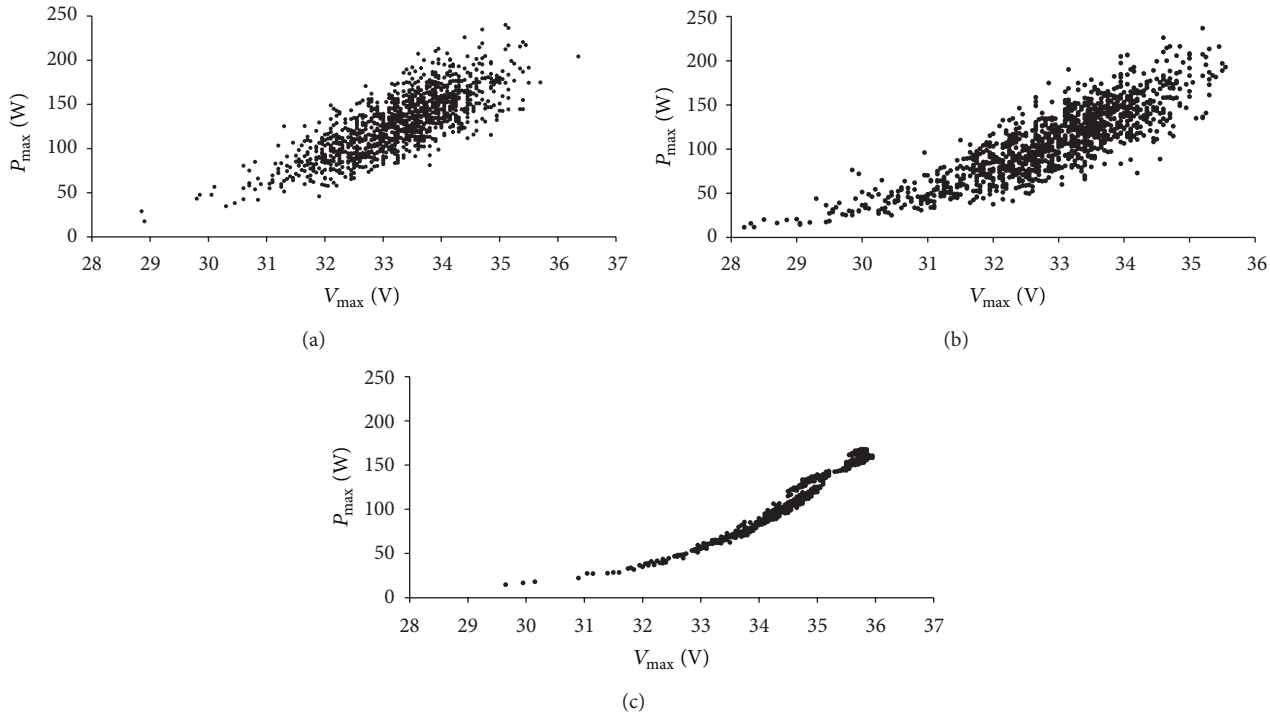

Figure 6: MPP distribution under simulated environment: (a) Gaussian distribution function generated environment data, (b) the set in (a) added with sun occultation by clouds, and (c) MPP distribution obtained using real weather data.

4. Configurations of Experiments

Experiments of MPPT of PV array utilizing the RLMPPT are conducted by simulation on computer and the results are compared with existing MPPT methods for PV array.

4.1. Environment Simulation and Configuration for PV Array. In this study, the PV array used for simulation is SY-175M, manufactured by Tangshan Shaiyang Solar Technology Co., Ltd., China. The power rating of the simulated PV array was 175 W with open-circuit voltage 44 V and short-circuit current 5.2 A. Validation of RLMPPT's effectiveness in MPPT was conducted via three experiments using two simulated and one real weather data sets. A basic experiment was first performed to determine whether the RLMPPT could achieve the task of MPPT by using a set of Gaussian distribution function generated temperature and irradiance. And the effect of sun occultation by clouds is added on the Gaussian distribution function generated set of temperature and irradiance as the second experiment. Real weather data for PV array, recorded in April, 2014, at Loyola Marymount University, California, USA, obtained from National Renewal Energy Laboratory (NREL) database provided an empirical data set to test the RLMPPT under real weather condition. Configurations of the experiments are described in the following.

Assume the PV array is located at the Subtropics area (Chiayi City, in the south of Taiwan) during 10:00 and 14:00 in summertime where the temperature and irradiance, respectively, are simulated using Gaussian distribution function with mean value of 30°C and 800 W/m^2 and standard deviation of 4°C and 50 W/m^2. According to (1), the simulated set of temperature and irradiance produces the MPP voltage (V_{mpp}) and the calculated MPP (P_{mpp}), as shown in Figure 6(a), where the V_{mpp} and the P_{mpp}, respectively, lay at x-axis and y-axis. Figures 6(b) and 6(c), respectively, show the (V_{mpp}, P_{mpp}) plots of the Gaussian generated temperature and irradiance with sun occultation effect and the real weather data recorded on April 1, 2014, at Loyola Marymount University.

4.2. Reward Function and State, Action Arrangement. In applying the RLMPPT to solve the problem of the MPPT for PV array, the reward function plays an important role, because not only a good reward function definition could achieve the right feedback on every execution of learning and tracking, but also it could enhance the efficiency of the learning algorithm. In this study, a hitting zone with *elliptical shape* for (V_{mpp}, P_{mpp}) is defined such that a positive reward value is given to the agent whenever it obtained (V_{mpp}, P_{mpp}) falls into the hitting zone in the sensing time slot. Figure 7 shows the different size of elliptical hitting zone superimposed on the plot of simulated (V_{mpp}, P_{mpp}) of Figure 6(a). The red elliptical circle in Figures 7(a), 7(b), and 7(c), respectively, represents that the hitting zone covers 37.2%, 68.5%, and 85.5% of the total (V_{mpp}, P_{mpp}) points, obtained from the simulated data of the previous day.

In realizing the RLMPPT, the state vector is defined as $S = [S_0, S_1, S_2, S_3] \subseteq \mathbf{S}$, while the meaning of each state is

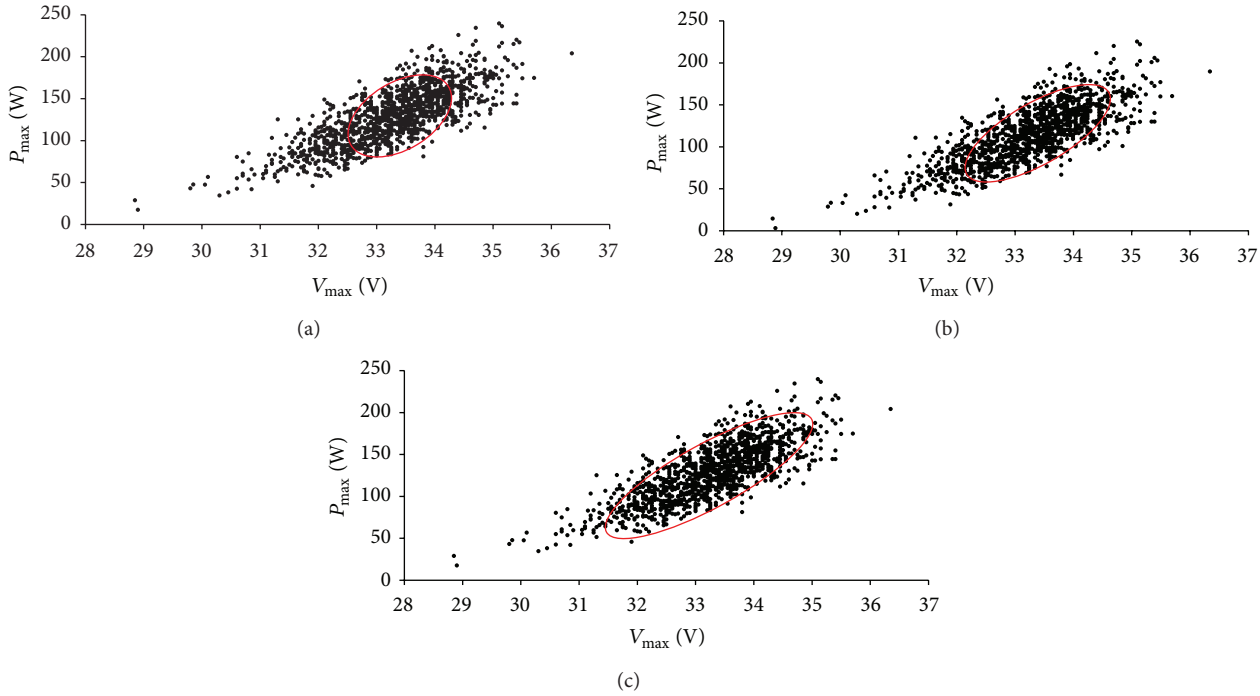

FIGURE 7: Different size of elliptical hitting zone superimposed on the plot of simulated (V_{mpp}, P_{mpp}) for the first set of simulated environment data: (a) 37.2% of the MPPs involved, (b) 68.5% of the MPPs involved, and (c) 85.5% of the MPPs involved.

explained in the previous section and shown in Figure 5. Six perturbations of $\Delta v(i)$ to the $V_{pv}(i)$ are defined as the set of action as follows:

$$\mathbf{A} = \{a_i \mid a_0 = -5\,\text{V},\ a_1 = -2\,\text{V},\ a_2 = -0.5\,\text{V},\ a_3 = +0.5\,\text{V},\ a_4 = +2\,\text{V},\ a_5 = +5\,\text{V}\}. \quad (15)$$

The ε-greedy is used in choosing agent's actions in the RLMPPT such that agent repeatedly selects the same action is prevented. The rewarding policy of the hitting zone reward function is to give a reward value of 10 and 0, respectively, to the agent whenever it obtained (V_{mpp}, P_{mpp}) in any sensing time slot falls-in and falls-out the hitting zone.

5. Experimental Results

5.1. Results of the Gaussian Distribution Function Generated Environment Data. In this study, experiments of RLMPPT in testing the Gaussian generated and real environment data are conducted and the results are compared with those of the P&O method and the open-circuit voltage method.

For the Gaussian distribution function generated environment simulation, the percentages of each RL agent choosing actions in early phase (0~25 minutes), middle phase (100~125 minutes), and final phase (215~240 minutes) of the RLMPPT are shown in the second, third, and fourth row, respectively, of Table 1 within two-hour period. It can be seen that, in the early phase of the simulation, the RL agent is in fast learning stage such that the action chosen by agent is concentrated on the ±5 V and ±2 V actions. However, in the middle and final phase of the simulation, the learning is completed and the agent exploited what it has learned; hence, the percentage of choosing fine tuning actions, that is, ±0.5 V, is increased from 20% of the early phase to 40% and 76%, respectively, for the middle and final phase. It can be concluded that the learning agent fast learned the strategy in selecting the appropriate action toward reaching the MPPT, and hence the goal of tracking the MPP is achieved by the RL agent.

TABLE 1: The percentage of choosing action in different phase.

Interval (minutes)	Action (V)					
	+5	+2	+0.5	−0.5	−2	−5
Early 0~25	16%	20%	8%	12%	24%	20%
Middle 100~125	4%	28%	24%	16%	20%	8%
Final 215~240	4%	8%	36%	40%	8%	4%

Experimental results of the offsets between the calculated and the tracking MPP by the RLMPPT and comparing methods at the sensing time are shown in Figure 8. Figures 8(a), 8(b), and 8(c), respectively, show the offsets between the calculated MPP and the tracking MPP by the P&O method, the open-circuit voltage method, and the RLMPPT method at the sensing time. One can see that, among the experimental results of the three comparing methods, the open-circuit voltage method obtained the largest offset, which is concentrated around 15 W. Even though the offsets obtained by the P&O method fall largely below 10 W, however, large portion of offset obtained by the P&O method also randomly scattered between 1 and 10 W. On the other hand, the proposed RLMPPT method achieves the least and condenses offsets

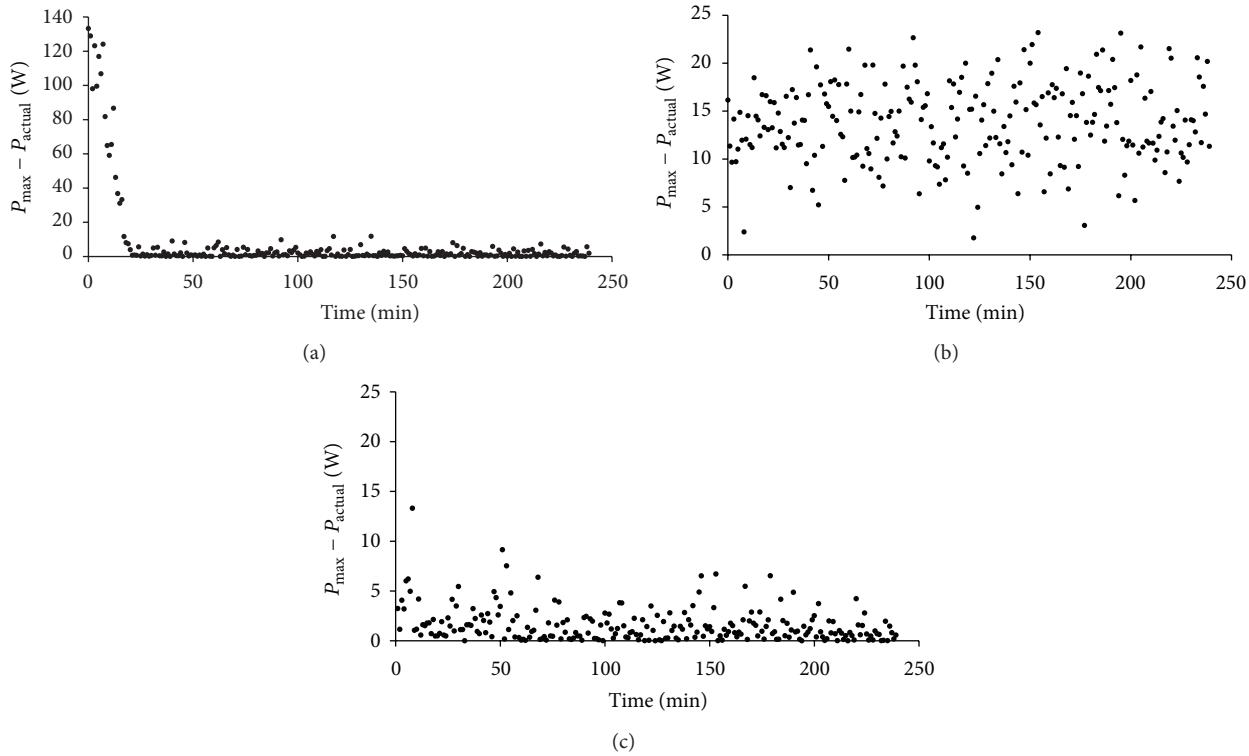

Figure 8: The maximum power point minus the predicted current power point corresponds to the same sensing time: (a) P&O method, (b) open-circuit voltage method, and (c) RLMPPT method.

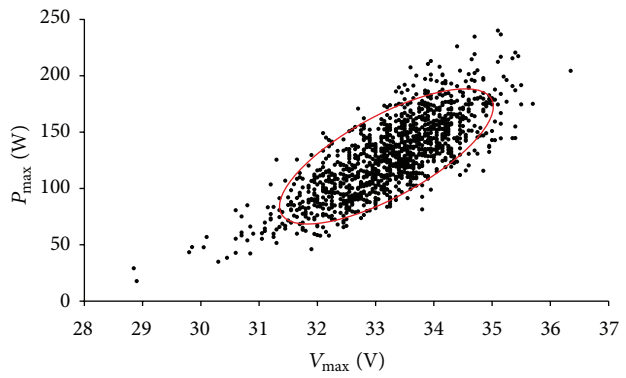

Figure 9: The definition of hitting zone for the 2nd set of experiment data.

below 5 W and only small portions fall outside of 5 W even in the early learning phase.

5.2. Results of the Gaussian Distribution Function Generated Environment Data with Sun Occultation by Clouds. In this experiment, the simulated data are obtained by adding a 30% chance of sun occultation by clouds to the test data in the first experiment such that the temperature will fall down 0 to 3°C and the irradiance will decrease 0 to 300 W/m^2. The experiment is conducted to illustrate the capability of RLMPPT in tracking the MPP of PV array under varied weather condition. Figure 9 shows the hitting zone definition

Table 2: The percentage of choosing action in different phase.

Interval (minutes)	Action (V)					
	+5	+2	+0.5	−0.5	−2	−5
Early 0~25	28%	8%	12%	8%	12%	32%
Middle 100~125	8%	16%	24%	20%	20%	12%
Final 215~240	4%	12%	40%	32%	8%	4%

for the simulated data with added sun occultation by clouds to the Gaussian distribution function generated environment data. The red elliptical shape in Figure 9 covers the 90.2% of the total (V_{mpp}, P_{mpp}) points, obtained from simulated data of the previous day. Table 2 shows the percentage of selecting action, that is, the perturbation $\Delta v(i)$ to the $V_{pv}(i)$, by the RLMPPT. The percentages of each RL agent choosing action in early phase, middle phase, and final phase of the RLMPPT method are shown in the second, third, and fourth row, respectively, of Table 2.

It can be seen from Table 2 that the percentages of selecting fine tuning actions, that is, the perturbation $\Delta v(i)$ is ±0.5 V, increase from 20% to 44%, and finally to 72%, respectively, for the early phase, middle phase, and final phase of MPP tracking via the RL agent. This table again illustrated the fact that the learning agent fast learned the strategy in selecting the appropriate action toward reaching the MPPT, and hence the goal of tracking the MPP of the PV array is achieved by the RL agent. However, due to the varied weather condition on sun occultation by clouds,

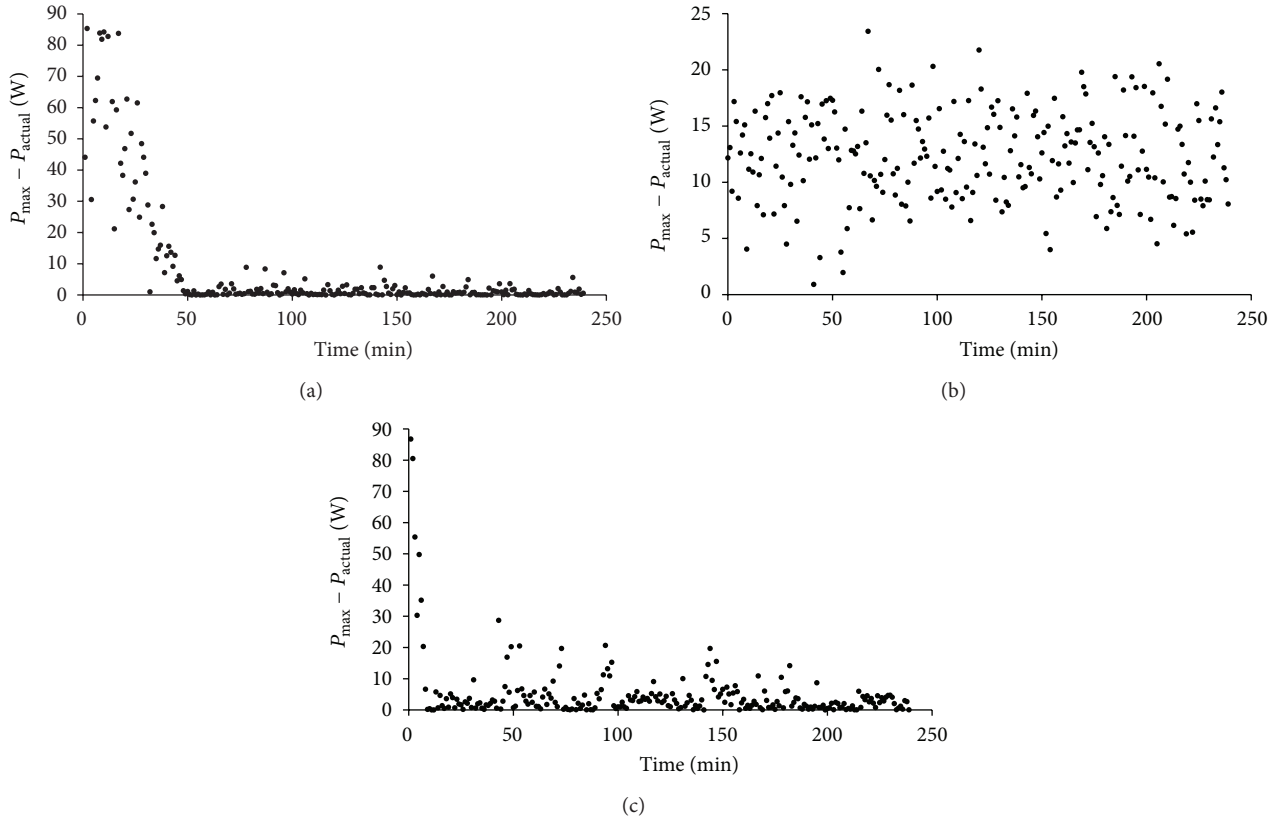

Figure 10: The maximum power point minus the predicted current power point corresponds to the same sensing time: (a) P&O method, (b) open-circuit voltage method, and (c) RLMPPT method.

the percentage of selecting fine-tuning actions is somewhat varied a little bit in comparison with the results obtained in Table 1, whose simulated data are generated by Gaussian distribution function without sun occultation effect.

Experimental results of the offsets between the calculated and the tracking MPP by the comparing methods of P&O, the open-circuit voltage, and the RLMPPT at the same sensing time are shown in Figures 10(a), 10(b), and 10(c), respectively. Experiment data from Figure 11 again exhibited that, among the three comparing methods, the open-circuit voltage method obtained the largest and sparsely distributed offsets data, which are concentrated around 15 W. Even though the offsets obtained by the P&O method fall largely between 0 and 5 W, however, large portion of offset obtained by the P&O method scattered around 1 to 40 W before the 50 minutes of the experiment. On the other hand, the proposed RLMPPT method achieves the least and condensed offsets below 5 W and mostly close to 3 W in the final tracking phase after 200 minutes.

5.3. Results of the Real Environment Data. Real weather data for PV array, recorded in April, 2014, at Loyola Marymount University, California, USA, is obtained online from National Renewal Energy Laboratory (NREL) database for testing the RLMPPT method under real environment data. The database is selected because the geographical location of the sensing station is also located in the subtropical area. A period of recorded data from 10:00 to 14:00 for 5 consecutive days is shown in Figure 6(c) and the hitting zone reward function for the data from earlier days is shown in Figure 11(a). The red elliptical shape in Figure 11(a) covers the 93.4% of the total (V_{mpp}, P_{mpp}) points, obtained from the previous 5 consecutive days of the NREL real weather data for testing. Figures 11(b) and 11(c), respectively, show the real temperature and irradiance data recorded at 04/01/2014 for generating the test data.

Table 3: The percentage of choosing action in different phase for the experiment with real weather data.

Interval (minutes)	Action (V)					
	+5	+2	+0.5	−0.5	−2	−5
Early 0~25	24%	16%	8%	12%	12%	28%
Middle 100~125	16%	16%	12%	24%	24%	8%
Final 215~240	4%	16%	32%	32%	12%	4%

Table 3 shows the percentage of selecting action by the RLMPPT. The percentages of each RL agent choosing action in early phase (0~25 minutes), middle phase (100~125 minutes), and final phase (215~240 minutes) of the RLMPPT method are shown in the second, third, and fourth rows, respectively, of Table 3.

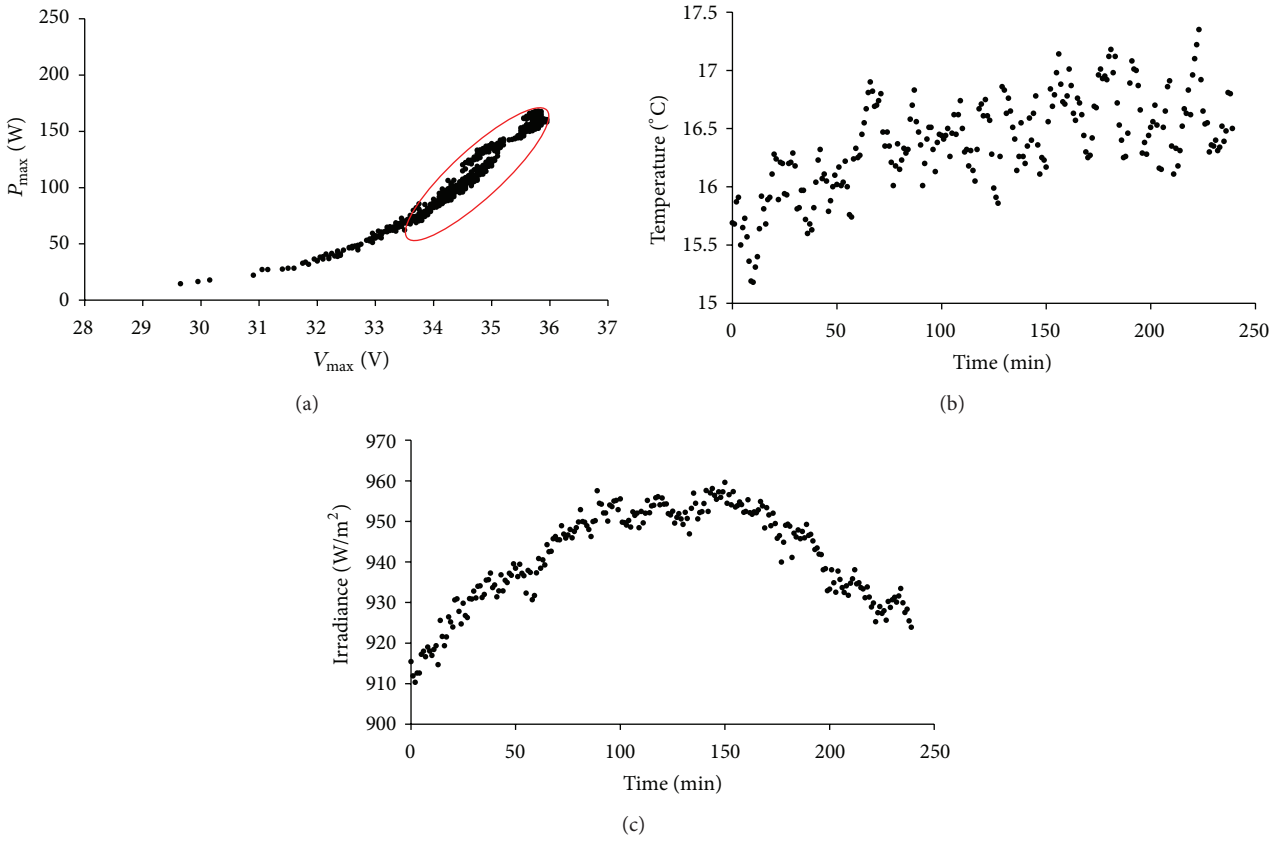

FIGURE 11: (a) The definition of hitting zone for the experiment using real weather data, (b) the real temperature data recorded in 04/01/2014 and used for generating the test data, and (c) the real irradiance data recorded at 04/01/2014 and used for generating the test data.

In Table 3, one can see that the percentage of selecting fine tuning actions, that is, the perturbation $\Delta v(i)$ is ± 0.5 V, increase from 20% to 36% and finally to 64%, respectively, for the early phase, middle phase, and final phase of MPP tracking via the RL agent. Even though the percentage of selecting fine tuning actions in this real data experiment has the least value among the three experiments, it exhibits that the RLMPPT learns to exercise the appropriate action of the perturbation $\Delta v(i)$ in tracking the MPP of the PV array under real weather data. This table again illustrated the fact that the learning agent fast learned the strategy in selecting the appropriate action toward reaching the MPP, and hence the goal of tracking the MPP of the PV array is achieved by the RL agent.

Experimental results of the offsets between the MPPT and the predicted MPPT by the comparing methods at the sensing time for the simulation data generated by the real weather data are shown in Figure 12. Figures 12(a), 12(b), and 12(c), respectively, show the offsets between the calculated MPPT and the tracking MPPT by the P&O method, the open-circuit voltage method, and the RLMPPT method. Experiment data from Figure 12 again shows that, among the three comparing methods, the open-circuit voltage method obtained the largest and sparsely distributed offsets data, whose distributions are concentrated within a band cover by two Gaussian distribution functions with the maximum offset value of 19.7 W. The offsets obtained by the P&O method fall largely around 5 and 2.5 W; however, large portion of offset obtained by the P&O method sharply decreased from 1 to 40 W in the early 70 minutes of the experiment. By the observation of Figure 12(c), the proposed RLMPPT method achieves the least and condenses offsets near 1 or 2 W and none of the offsets is higher than 5 W after 5 simulation minutes from the beginning.

5.4. Performance Comparison with Efficiency Factor. In order to validate whether the RLMPPT method is effective or not in tracking MPP of a PV array, the efficiency factor η is used to compare the performance of other existing methods for the three experiments. The η *is* defined as follows:

$$\eta_{\text{mppt}} = \frac{\int_0^t P_{\text{actual}}(t)\,dt}{\int_0^t P_{\text{max}}(t)\,dt}, \tag{16}$$

where $P_{\text{actual}}(t)$ and $P_{\text{max}}(t)$, respectively, represent the tracking MPP of the MPPT method and the calculated MPP. Table 4 shows that, for the three test data sets, the open-circuit voltage method has the least efficiency factor among the three comparing methods, and the RLMPPT method has the best efficiency factor which is slightly better than that of P&O method. The advantage of the RLMPPT method over the P&O method is shown in Figures 10 and 12 where not only

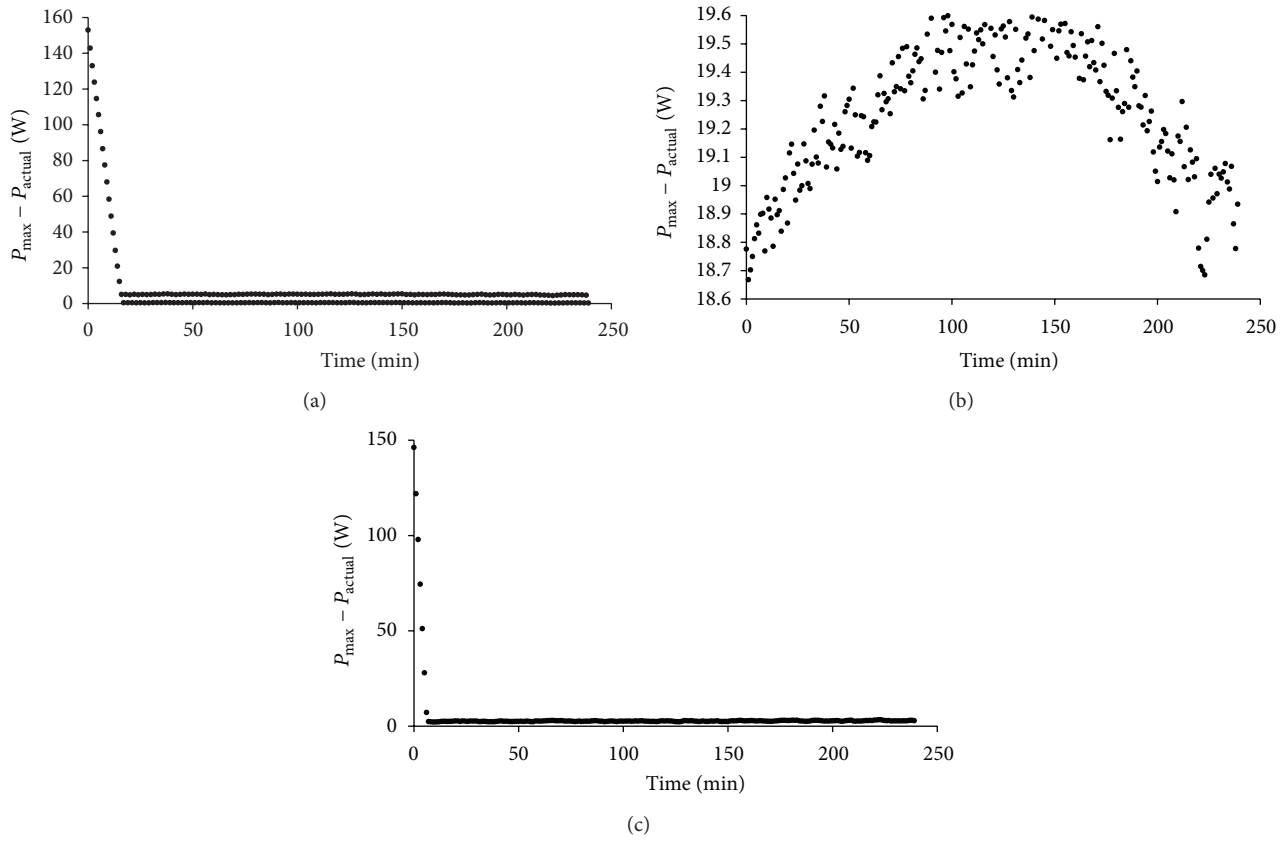

Figure 12: The maximum power point minus the predicted current power point corresponds to the same sensing time: (a) P&O method, (b) open-circuit voltage method, and (c) RLMPPT method.

Table 4: Comparison of efficiency factor for the three comparing methods.

Learning phase (min)	Methods		
	Open-circuit voltage	P&O	RLMPPT
Early 10:00~10:25	86.9%	83.5%	89.2%
Middle 11:40~12:05	87.5%	97.8%	99.4%
Final 13:35~14:00	87.8%	97.7%	99.3%
Overall 10:00~14:00	87.6%	95.4%	98.5%

the RLMPPT method significantly improves the efficiency factor in comparing that of the P&O, but also the learning agent of the RLMPPT fast learned the strategy in selecting the appropriate action toward reaching the MPPT which is much faster than the P&O method in exhibiting a slow adaptive phase. Hence, the RLMPPT not only improves the efficiency factor in tracking the MPP of the PV array, but also has the fast learning capability in achieving the task of MPPT of the PV array.

6. Conclusions

In this study, a reinforcement learning-based maximum power point tracking (RLMPPT) method is proposed for PV array. The RLMPPT method monitors the environmental state of the PV array and adjusts the perturbation to the operating voltage of the PV array in achieving the best MPP. Simulations of the proposed RLMPPT for a PV array are conducted on three kinds of data set, which are simulated Gaussian weather data, simulated Gaussian weather data added with sun occultation effect, and real weather data from NREL database. Experimental results demonstrate that, in comparison to the existing P&O method, the RLMPPT not only achieves better efficiency factor for both simulated and real weather data sets but also adapts to the environment much fast with very short learning time. Further, the reinforcement learning-based MPPT method would be employed in the real PV array to validate the effectiveness of the proposed novel MPPT method.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] T. Esram and P. L. Chapman, "Comparison of photovoltaic array maximum power point tracking techniques," *IEEE Transactions on Energy Conversion*, vol. 22, no. 2, pp. 439–449, 2007.

[2] D. P. Hohm and M. E. Ropp, "Comparative study of maximum power point tracking algorithms," *Progress in Photovoltaics: Research and Applications*, vol. 11, no. 1, pp. 47–62, 2003.

[3] N. Femia, G. Petrone, G. Spagnuolo, and M. Vitelli, "Optimization of perturb and observe maximum power point tracking method," *IEEE Transactions on Power Electronics*, vol. 20, no. 4, pp. 963–973, 2005.

[4] J. S. Kumari, D. C. S. Babu, and A. K. Babu, "Design and analysis of PO and IPO MPPT technique for photovoltaic system," *International Journal of Modern Engineering Research*, vol. 2, no. 4, pp. 2174–2180, 2012.

[5] D. P. Hohm and M. E. Ropp, "Comparative study of maximum power point tracking algorithms using an experimental, programmable, maximum power point tracking test bed," in *Proceedings of the Conference Record of the IEEE 28th Photovoltaic Specialists Conference*, pp. 1699–1702, 2000.

[6] L.-R. Chen, C.-H. Tsai, Y.-L. Lin, and Y.-S. Lai, "A biological swarm chasing algorithm for tracking the PV maximum power point," *IEEE Transactions on Energy Conversion*, vol. 25, no. 2, pp. 484–493, 2010.

[7] T. L. Kottas, Y. S. Boutalis, and A. D. Karlis, "New maximum power point tracker for PV arrays using fuzzy controller in close cooperation with fuzzy cognitive networks," *IEEE Transactions on Energy Conversion*, vol. 21, no. 3, pp. 793–803, 2006.

[8] N. Mutoh, M. Ohno, and T. Inoue, "A method for MPPT control while searching for parameters corresponding to weather conditions for PV generation systems," *IEEE Transactions on Industrial Electronics*, vol. 53, no. 4, pp. 1055–1065, 2006.

[9] G. C. Hsieh, H. I. Hsieh, C. Y. Tsai, and C. H. Wang, "Photovoltaic power-increment-aided incremental-conductance MPPT with two-phased tracking," *IEEE Transactions on Power Electronics*, vol. 28, no. 6, pp. 2895–2911, 2013.

[10] R. L. Mueller, M. T. Wallace, and P. Iles, "Scaling nominal solar cell impedances for array design," in *Proceedings of the IEEE 1st World Conference on Photovoltaic Energy Conversion*, vol. 2, pp. 2034–2037, December 1994.

[11] N. Femia, G. Petrone, G. Spagnuolo, and M. Vitelli, "Optimizing sampling rate of P&O MPPT technique," in *Proceedings of the IEEE 35th Annual Power Electronics Specialists Conference (PESC '04)*, vol. 3, pp. 1945–1949, June 2004.

[12] N. Femia, G. Petrone, G. Spagnuolo, and M. Vitelli, "A technique for improving P&O MPPT performances of double-stage grid-connected photovoltaic systems," *IEEE Transactions on Industrial Electronics*, vol. 56, no. 11, pp. 4473–4482, 2009.

[13] K. H. Hussein, I. Muta, T. Hoshino, and M. Osakada, "Maximum photovoltaic power tracking: an algorithm for rapidly changing atmospheric conditions," *IEE Proceedings—Generation, Transmission and Distribution,*, vol. 142, no. 1, pp. 59–64, 1995.

[14] C. Hua, J. Lin, and C. Shen, "Implementation of a DSP-controlled photovoltaic system with peak power tracking," *IEEE Transactions on Industrial Electronics*, vol. 45, no. 1, pp. 99–107, 1998.

[15] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[16] A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.

[17] C. J. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.