Journal of Systems Engineering and Electronics
Vol. 29, No. 6, December 2018, pp.1293 – 1299

UAV flight strategy algorithm based on dynamic programming

ZHANG Zixuan[1], WU Qinhao[2], ZHANG Bo[3,*], YI Xiaodong[1], and TANG Yuhua[1]

1. College of Computer, National University of Defense Technology, Changsha 410073, China;
2. College of Electronic Science, National University of Defense Technology, Changsha 410073, China;
3. National Innovative Institute of Defense Technology, Beijing 100010, China

Abstract: Unmanned aerial vehicles (UAVs) may play an important role in data collection and offloading in vast areas deploying wireless sensor networks, and the UAV's action strategy has a vital influence on achieving applicability and computational complexity. Dynamic programming (DP) has a good application in the path planning of UAV, but there are problems in the applicability of special terrain environment and the complexity of the algorithm. Based on the analysis of DP, this paper proposes a hierarchical directional DP (DDP) algorithm based on direction determination and hierarchical model. We compare our methods with Q-learning and DP algorithm by experiments, and the results show that our method can improve the terrain applicability, meanwhile greatly reduce the computational complexity.

Keywords: motion state space, map stratification, computational complexity, dynamic programming (DP), environmental adaptability.

DOI: 10.21629/JSEE.2018.06.16

1. Introduction

The use of UAVs has become more widespread. In particular, advances in sensors and communications equipment have led to the development and large-scale use of UAVs in recent years [1], especially in the areas of polar climate monitoring [2], providing communication security in the disaster area [3] and the management of forest fire hazards in mountainous areas [4]. However, the energy modules that UAV can carry are limited to volume [5]. They tend to detect selective parts of the region because of the battery life, which is even more obvious when the target range is larger. In the process of reconnaissance, the relationship between energy, efficiency and communication quality becomes an important issue.

The current UAV communication field mainly focuses on the signal transmission scheme, the energy optimization and relay deployment strategy [6 – 9]. Attention has also been paid to the impact of action strategy optimization of the UAV/UGV in a large area with sparsely distributed sensor networks at points of interest (POIs) [10,11]. However, the related works consider the case that the terrain poses no constraints on the traveling path, which is not realistic.

The problem to be solved in this paper is explained as follows: We consider such a scene in which the target detection area is a vast area with a sparse distribution of POIs and unreachable locations. Only in the designated area the UAVs can establish reliable communication with the data aggregation centre. The UAV needs to detect as many POIs as possible with limited energy, which results in the limited moving capability of UAVs. The problem is converted to a UAV starting from the designated point of the target area. By the strategy search, the communication link with the data aggregation centre is established and the data offloading is completed.

The problem is inherently a multi-step optimization problem and the dynamic programming (DP) algorithm can be used to solve the problem. However, as analysed in the rest of the paper, the DP algorithm may not directly handle the black points (unreachable locations) and also incurs high computational complexity. Therefore, we propose an optimized algorithm based on direction determination called directional dynamic programming (DDP) algorithm and hierarchical modeling, which is suitable for general terrain conditions and improves the applicability of the algorithm. The computational efficiency of the algorithm in a sparse-distributed environment is improved, and the time complexity on similar problems can be effectively reduced. Q-learning is a very important algorithm in rein-

Manuscript received April 17, 2018.

*Corresponding author.

This work was supported by the National Natural Science Foundation of China (91648204; 61601486), State Key Laboratory of High Performance Computing Project Fund (1502-02), and Research Programs of National University of Defense Technology (ZDYYJCYJ140601).

forcement learning, which is widely applied in UAV path planning [12,13]. It has obvious advantages in terms of algorithm applicability and efficiency. We use the Q-learning algorithm as a benchmark to compare the performance of the DDP algorithm.

The rest of the paper is organized as follows: In Section 2, we first propose the problem formulation, then introduce the hierarchical DDP algorithm in detail. In Section 3, the commonly used Q-learning algorithm and the original DP algorithm are compared by numerical experiments, which proves that the hierarchical DDP algorithm has a great advantage in reducing computational complexity and a superiority in applications of complex terrain. Finally, our future work is explained in Section 4.

2. Modeling and algorithm design

2.1 Subsection gridding of the map

Referring to the existing modeling approach, we consider the whole area is gridded and the specific reconnaissance gain is the quantitative value in each grid. This quantification is modeled on the underlying wireless sensor network (WSN) topology and the amount of perceived information. In this paper, we consider the general situation where the gains are not evenly distributed in the region. Then we consider that the UAV accesses the grid of multiple WSNs and collects data. It eventually reaches the destination, establishes a communication link, and offloads the perceived data.

2.2 Ideas and problems of traditional DP algorithm

In traditional DP, the algorithm meshes the task area, and each grid (i.e., the location of the UAV) is a state in the algorithm state space. In order to explicitly specify the difference, the state space generated by this modeling method is called the location grid state. Through the iteration of each state, the optimal path is obtained by ascending the iterative results. There are the following problems in this algorithm:

(i) Repeated access: The algorithm cannot realize repeated access to the same grid, since the final result is arranged in an ascending order of iterations of multiple states. In fact, the UAV may require multiple reconnaissance to get comprehensive information.

(ii) Black points (BPs): The algorithm is not adaptable under special terrain conditions. We define the areas that the UAVs cannot enter into as BP, e.g., surrounded by unreachable locations like cave or valley. The UAV will stay in the grid which represents the end of the cave because DP does not allow it to repeatedly access the previous grid, which is the only way out.

(iii) Complexity: The computational complexity is prohibited. The traditional DP algorithm will iterate the whole reconnaissance area indistinguishably. In order to achieve a certain precision, the algorithm will divide the whole area into a high order gridding. The result is that the algorithm has a disadvantage in computing complexity. In fact, we only care about some of the key points in vast area, defined as POI.

In the rest of this section, we first design new modeling methods and DDP algorithm to deal with the repeated access and BP problems. Then we introduce hierarchical modeling to exploit sparsity in reducing computing complexity.

2.3 Design and modeling of DDP algorithm

In traditional DP, a grid is regarded as a possible state, and the motion is modeled as the transfer between states. In comparison, in DDP we model the transfer from an edge to an edge as a motion state and 16 motion states may exist for a single grid, as shown in Fig. 1.

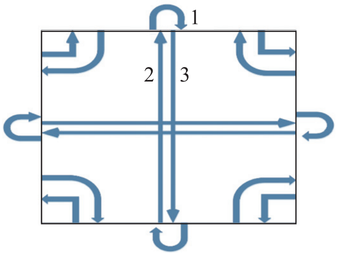
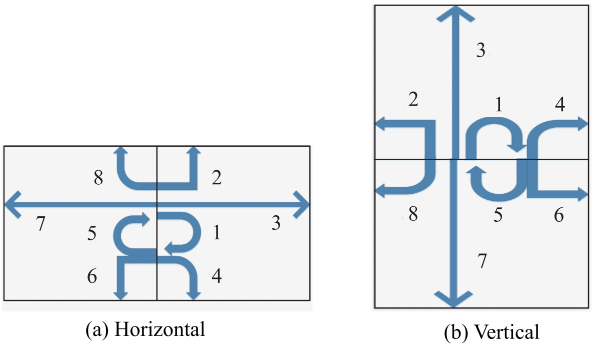


Fig. 1 Illustration of 16 motion states in a single grid

On this basis, all logic partitioning methods of the motion state are as follows in Fig. 2. It corresponds to the next possible motion state of the UAVs at the vertical and the horizontal edges.



(a) Horizontal (b) Vertical

Fig. 2 Motion states of horizontal and vertical edges

For a 4×4 grid, there are 40 edges and 320 motion states, where logical links constitute the entire state space. In this problem, we assume that a UAV is moving in logical space. Each moving step takes a time step i. The state space s is expressed as

$$\{(x_1, y_1, m, n), (x_2, y_2, m, n), (x_3, y_3, m, n), \ldots\}$$

where x and y represent the coordinates of the edges, while $m \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ represents the eight motion states on the edge. $n \in \{1, 2\}$ represents the horizontal or vertical edges respectively. The action space is $u \in \{1, 2, 3, 4, 5, 6, 7, 8\}$, which represents the next eight motion states that the next step of each edge can take.

The gain function is denoted by $g(s)$, and its value corresponds to 320 motion states and function. $f(s, u)$ denotes the state transition function, which is the new motion state that the UAV can arrive after taking action u in the state of s.

The value function is defined as $J(s)$ which can be defined as the future gains got by taking a set of policy π starting from a determined single state. The initial value of $J(s)$ is defined as 0. That is $J^0(s) = 0$, for all s. T is defined as

$$J^{k+1}(s) = TJ^k(s) = \max_u[\gamma J^k(f(s, u)) + g(s)] \quad (1)$$

where γ is called discount factor to balance the present and future gains ($\gamma < 1$). The final results are the final maximum value of $J^*(s)$ after multiple iterations. From the final value of $J^*(s)$ for each state s_i, we can get the sequence of the policy π. The ascending order is the final result. Taking the upper limit of the iteration as β, the iteration termination condition is

$$\max(J^{k+1}(s) - J^k(s)) \leqslant \beta, \text{ for all } s. \quad (2)$$

The action u_i of each time step i is defined by

$$u_i = \pi^*(s_i) = \arg\max_u[\gamma J^*(g(s_i) + f(s_i, u))]. \quad (3)$$

It is possible to arrive at a more realistic result. These states contain repetitive access to a single state in the original DP state space, as shown in state 1 of Fig. 1. It distinguishes the direction of the two states when switching, such as states 2 and 3.

The modeling change means that the state space changes from the movement between grids of DP to the movement between edges in the proposed DDP. However, all the states of the J value are only taken once in the process of generating the policy. This operation is to form a cycle when the motion state 1 or state 5 is adopted, which allows repeatedly traversing a grid. More importantly, the above modeling allows the UAV to get out of a grid surrounded by BPs.

2.4 Computational complexity optimization for sparse environments

In realistic problems, the target distribution in the reconnaissance area is often sparse. The locations of the POIs are sparse and partially concentrated, and the distribution is uneven, which is defined as “sparse environment” for convenience. We consider the use of multi-layer structure of the hierarchical method to solve the problem with large-scale computational complexity. The DDP algorithm serves as the top-level algorithm, which is defined as “map layer algorithm”. The exhaustion method is used at the underlying algorithm, which is defined as “area layer algorithm” to provide the required data to the upper layer. The underlying algorithm relies on the detailed WSN topology in a grid to quantify the value of the POIs. The overall algorithm structure diagram is shown as follows in Fig. 3.

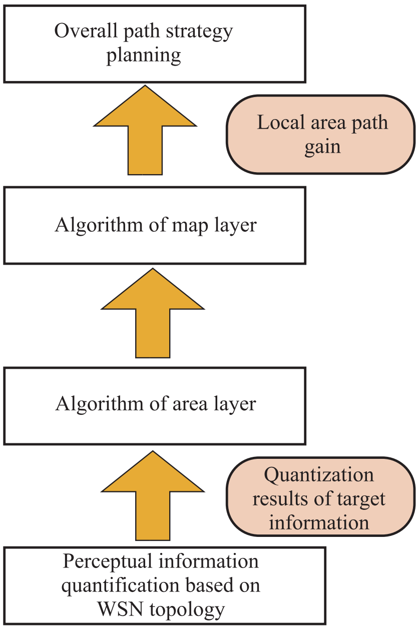

Overall path strategy planning
Local area path gain
Algorithm of map layer
Algorithm of area layer
Quantization results of target information
Perceptual information quantification based on WSN topology

Fig. 3 Overall algorithm structure diagram

In the top-level DDP algorithm, the $g(s)$ function values of the 16 motion states for each grid are provided by the underlying algorithm. The modeling method is to further divide the individual grid of the DDP into K cells ($K = 16$ in our experiment) where each cell is given a gain value. We can pre-specify four imports and exports in a single grid. By specifying a fixed number of step i, the algorithm will exhaust all the paths that match the import and export settings and add the gain values for all paths.

For each combination of the imports and exports, the gain on all the points on the path is calculated only once, and the maximum total gain is chosen as the gain of this movement at the top-level. The advantage of this approach is that the hierarchical method only needs to focus on POIs in a wide range of sparse environments, which is sparsely distributed. In this paper, we use the exhaustion method as the underlying algorithm to present our initial findings. However, the complexity of the exhaustion method may be replaced by more advanced modules having better computational efficiency in various scenarios.

3. Experiment and verification

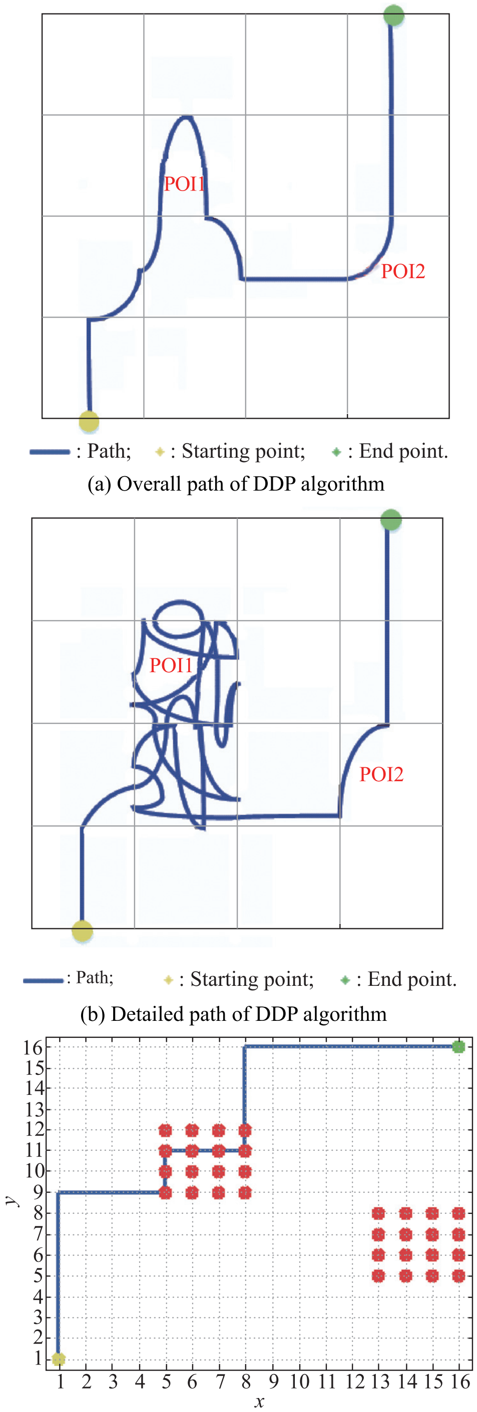
In this section, we conduct three groups of experiments. In Section 3.1, we test the performance of the hierarchical DDP algorithm in general. Under the same conditions, we compare the results with the results of Q-learning and DP algorithms. We can conclude that hierarchical DDP can carry out multi-path reconnaissance in key areas to obtain higher gains. In Section 3.2, we expand the algorithm scale and compare the calculation time of DP and hierarchical DDP algorithms. The hierarchical DDP algorithm greatly improves the computational complexity, especially when the problem scale is further expanded. In Section 3.3, we test the performance of three algorithms when BPs are distributed to valley terrain. As a result, the DP algorithm cannot handle this problem, and the UAV stops at the end of the valley. The Q-learning algorithm fails to go deep into the valley end. Hierarchical DDP can solve this problem, which makes the algorithm more practical.

3.1 Analysis of results with hierarchical DDP algorithm

We use the pre-set data set to test the hierarchical DDP algorithm. Consider the DP algorithm (4×4 scale) in upper layer and the lower layer algorithm (4×4 scale) respectively.

The experimental conditions are set as follows: the discount factor $\gamma = 0.9$, the iterative threshold is 0.5, and the number of steps in the underlying algorithm is limited to seven steps, and two POIs are set, called POI 1 and POI 2. The parameters in Q-learning algorithm are as follows: the maximum number of iterations is 1 000, while the greediness is 0.9, the discount factor $\gamma = 0.9$ and the learning rate is 0.5.

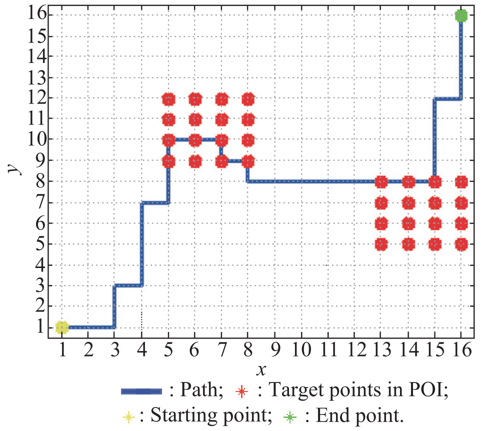
The optimized paths of each algorithm are given by Fig. 4. Fig. 4(a) is a holistic path diagram, which reflects the transfer between grids. However, the UAV may detect a grid in a variety of ways, where Fig. 4(b) is a specific step within grids for Fig. 4(a). The UAV will search for multiple paths in POI 1 and its surrounding grids do not adopt repeated motions. In the rest of this paper, only the holistic path diagram of DDP algorithm is provided in the following results for the sake of simplicity. For the DDP algorithm, the optimal path is often more than one, so only one of them is given in order to simplify the process. Even if the specific path is different, the algorithm gives the same gain for all the optimized paths. Since DP adopts the new algorithm modeling design, the state space changes from point to motion. Therefore, in the presentation of experimental results, the DDP algorithm is very different from the DP and Q-learning algorithms. DDP uses map layering design, so when the result of the path is presented, it mainly displays the path in the upper grid. The other two algorithms are presented directly on the underlying grid, so the result graph is different. In fact, the experimental conditions are exactly the same. The result of the DP algorithm is shown in Fig. 4(c).



(a) Overall path of DDP algorithm

(b) Detailed path of DDP algorithm

(c) Path of DP algorithm



(d) Path of Q-learning algorithm

Fig. 4 Simulated results with hierarchical DDP algorithm, DP algorithm and Q-learning algorithm

The path in the DP algorithm directly passes through the POI 1, and fails to get more gains. At the same time, the path in the DP algorithm also gives up the access to POI 2. The result of the Q-learning algorithm is shown in Fig. 4(d). The path in the algorithm also fails to get more gains at the POI locations.

From the experimental results, we can see that the DDP algorithm performs better on general problems compared with the other two algorithms. The DDP algorithm can take different paths in the POI locations for repeated paths, improve the reconnaissance gains, and will not give up the detection of other POIs.

3.2 Comparison of complexity

On the basis of the previous section, we add two BPs and increase the total number of the grids. We compare the DP algorithm and analyse the results. The DP algorithm needs to deal with all the expansion of the data set which means the states 16×16, a total of 256 states. The BPs in the hierarchical DDP algorithm are represented by 16 consecutive BPs points in DP.

In order to further verify the conclusion, we continue to expand the dimension of the algorithm, and the underlying algorithm keeps the grid segmentation of 4×4. We test the performance of hierarchical DDP in the scale of 5×5, 7×7, 9×9, 11×11 and corresponding DP in 20×20, 28×28, 36×36, 44×44 respectively. The computation time of the algorithm is shown in Fig. 5.

As can be seen from the results, the computation time of the hierarchical DDP algorithm is much shorter than the DP algorithm and the scalability of hierarchical DDP is much better than DP. This proves that hierarchical DDP is superior in computational complexity in sparse distributed environments. In this section, no experimental results are compared with the Q-learning algorithm. There are three main reasons. First, because of the greedy decision, the Q-learning algorithm is not monotonically decreasing, and it may not have convergence, so we may not strictly compare the computation time. Second, the computation time of Q-learning algorithm depends largely on the number of iterations. The choice of the number of iterations depends on the need for the answer rather than the algorithm itself. Third, Q-learning does not apply to large-scale algorithms. When the dimensions of the problem are relatively high, the computation efficiency of the Q-learning algorithm is relatively low at the beginning of the exploratory stage.

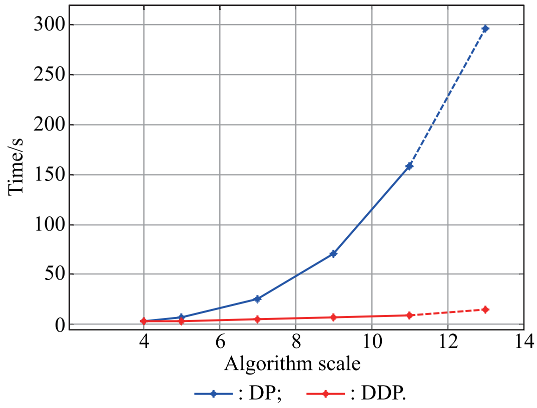
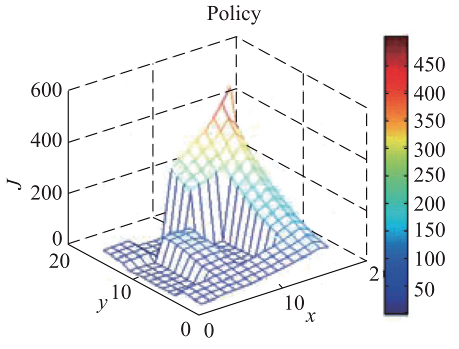


Fig. 5 Computation time of DP and hierarchical DDP

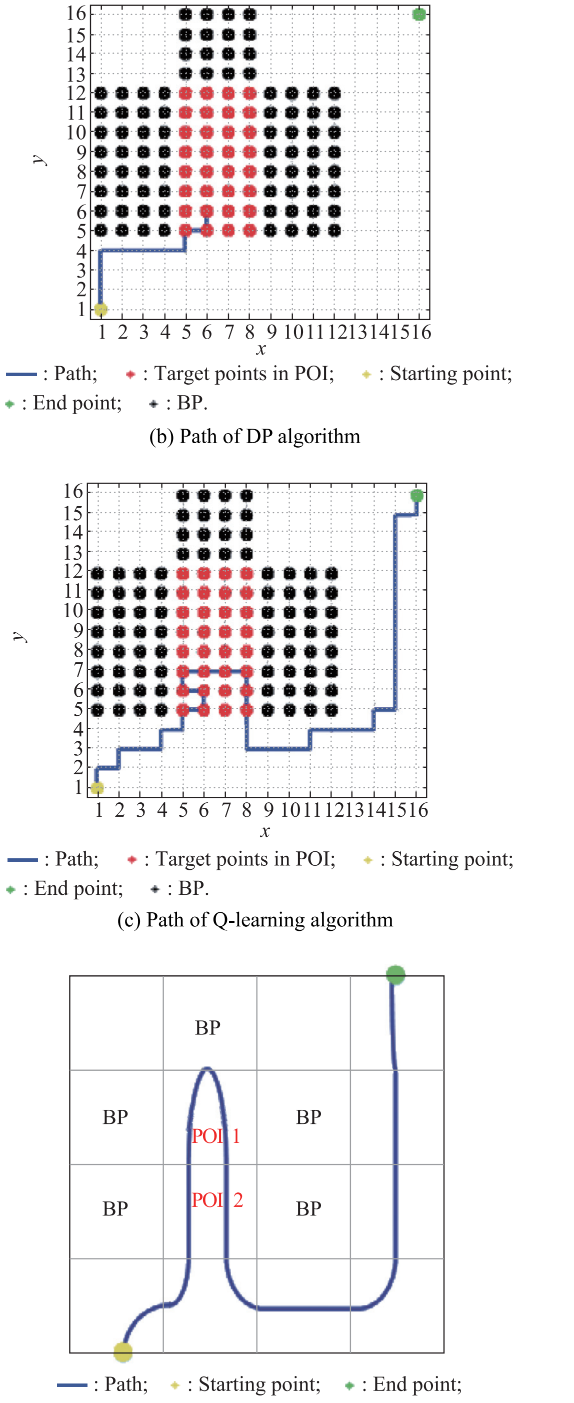
3.3 Comparison of algorithm applicability

Now consider the applicability of the algorithm under special terrain conditions where we form a terrain like a valley. The experimental set up is consistent with Section 3.1. DP algorithm results are shown in Fig. 6(a). The path map generated by the DP algorithm is shown in Fig. 6 (b). The path graph generated by the Q-learning algorithm is shown in Fig. 6 (c). The hierarchical DDP algorithm results are shown in Fig. 6(d).

When the DP algorithm is applied, the UAV will be locked in the POI 1 position (in the valley) and cannot be withdrawn due to the ascending order of J, when entering the valley to detect POI 1.



(a) J function of DP algorithm



(b) Path of DP algorithm

(c) Path of Q-learning algorithm

(d) Overall path of DDP algorithm

Fig. 6 Comparison of results with DP algorithm, hierarchical DDP algorithm and Q-learning algorithm under valley terrain

The path in the Q-learning algorithm cannot go into the valley to detect it. It will fly over the entrance of the valley rather than into the valley. The hierarchical DDP algorithm, allowing repeated access of a grid, is capable of supporting planning in such terrains. Compared with Q-learning, DDP can achieve better reconnaissance effect. Therefore, it may be noted that the proposed hierarchical DDP algorithm has better applicability under various terrain conditions.

4. Conclusions

This paper considers the problem of UAV planning in vast areas with a sparse distribution of POIs and unreachable locations. Conventional DP based on location grid states is a natural choice. Since it does not deal with the repeated access and BP problems, and its complexity is high and not scalable, the DP algorithm is not applicable to the problems that we want to solve in this paper. Hence we propose the DDP algorithm and the hierarchical method to solve the problem. In the experiment, we compare the DDP algorithm with the traditional DP algorithm and the Q-learning algorithm, which is widely used in reinforcement learning. From the experimental results, the hierarchical DDP algorithm improves the applicable field of the algorithm by solving the problems of repeated access and BP in the DP algorithm, while also greatly improves the computational efficiency. Compared with the Q-learning algorithm, the DDP algorithm also improves the state space and the movement space, and improves the reconnaissance gain of the algorithm around the target. Furthermore, the algorithm we propose has a certain application value and prospect in the path planning of such platforms as robots, UGVs and AUVs.

References

[1] ABDULLA A E, FADLULLAH Z M, NISHIYAMA H, et al. An optimal data collection technique for improved utility in UAS-aided networks. Proc. of the IEEE Conference on Computer Communications, 2014: 736 – 744.

[2] CURRY J A, MASLANIK J, HOLLAND G, et al. Applications of aerosondes in the arctic. Bulletin of the American Meteorological Society, 2004, 85(12): 1855 – 1861.

[3] TAKAHASHI Y, KAWAMOTO Y, NISHIYAMA H, et al. A novel radio resource optimization method for relay-based unmanned aerial vehicles. IEEE Trans. on Wireless Communications, 2018, 17(11): 7352 – 7363.

[4] FREW E W, BROWN T X. Airborne communication networks for small unmanned aircraft systems. Proceedings of the IEEE, 2008, 96(12): 2008 – 2027.

[5] ABDULLA A E, NISHIYAMA H, YANG J, et al. Hymn: a novel hybrid multi-hop routing algorithm to improve the longevity of WSNs. IEEE Trans. on Wireless Communications, 2012, 11(7): 2531 – 2541.

[6] ABDULLA A E, FADLULLAH Z M, NISHIYAMA H, et al. Toward fair maximization of energy efficiency in multiple UAS-aided networks: a game-theoretic methodology. IEEE Trans. on Wireless Communications, 2015, 14(1): 305 – 316.

[7] FADLULLAH Z M, TAKAISHI D, NISHIYAMA H, et al. A dynamic trajectory control algorithm for improving the communication throughput and delay in UAV-aided networks. IEEE Network, 2016, 30(1): 100 – 105.

[8] LI K, NI W, WANG X, et al. Energy-efficient cooperative relaying for unmanned aerial vehicles. IEEE Trans. on Mobile Computing, 2016, 15(6): 1377 – 1386.

[9] FOTOUHI A, DING M, HASSAN M. Dynamic base station repositioning to improve spectral efficiency of drone small cells. arXiv preprint arXiv: 2017, 1704.01244.

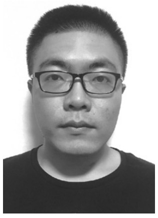
[10] XIE L, SHI Y, HOU Y T, et al. On renewable sensor networks with wireless energy transfer: the multi-node case. Proc. of the 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, 2012: 10 – 18.

[11] XIE L, SHI Y, HOU Y T, et al. A mobile platform for wireless charging and data collection in sensor networks. IEEE Journal on Selected Areas in Communications, 2015, 33(8): 1521 – 1533.

[12] FERRARI S, DAUGHERTY G. Q-learning approach to automated unmanned air vehicle (UAV) demining. Journal of Defense Modeling & Simulation, 2010(9): 76920S-76920S-12.

[13] ZHAO Y Y, ZHENG Z, ZHANG X Y, et al. Q learning algorithm based UAV path learning and obstacle avoidence approach. Proc. of the 36th Chinese Contrd Conference, 2017: 3397 – 3402.

Biographies



ZHANG Zixuan was born in 1993. He received his B.S. degree in College of Electronic Science from National University of Defense and Technology (NUDT), Changsha, China, in June 2016. Now, he is a master of College of Computer from NUDT. His current research field includes MIMO technology, satellite communication and joint motion planning for UAV communication.
E-mail: zx.zhang16@hotmail.com

WU Qinhao was born in 1993. Now, he is a master of College of Electronic Science and Engineering from National University of Defense and Technology. His current research field includes metamaterial antenna, radar coincidence imaging and radar signal processing.
E-mail: qinhaowu@hotmail.com

ZHANG Bo was born in 1989. He is a Ph.D. and currently an assistant professor in Artificial Intelligence Research Center (AIRC), National Innovative Institute of Defense Technology. His research interests in wireless communications include the design and analysis of cooperative communications, multiple-input-multiple-output systems, and network-robotic systems.
E-mail: zhangbo10@nudt.edu.cn

YI Xiaodong was born in 1978. He is a Ph.D. and currently a researcher in the State Key Laboratory of High Performance Computing. His research interests are operating system, high performance computing, robotics software, etc.
E-mail: yixiaodong@nudt.edu.cn

TANG Yuhua was born in 1962. She is currently a professor in the State Key Laboratory of High Performance Computing, National University of Defense Technology. Her research interests include supercomputer architecture and core router's design.
E-mail: yhtang@nudt.edu.cn