Route Optimization of Unmanned Aerial Vehicle by using Reinforcement Learning

M Kundu* and D J Nagendra Kumar

1Department of Computer Science and Engineering, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India-534202
2Department of Information and Technology, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India-534202

*E-mail: kundu.mayuri@gmail.com

Abstract: The study proposes the machine learning based algorithm for autonomous vehicle. The dynamic characteristic of unmanned aerial vehicle and real time disturbances such as wind currents, obstacles are considered. The novelty of the work lies in the introduction of reinforcement learning for achieving optimized path which can be followed by the unmanned aerial vehicles to complete the tour from initial to destination point. The feasible optimal route may be found by incorporating the algorithm with a reasonable optimality and computational cost when map and current field data are given.

Keywords — Unmanned aerial vehicle, Route optimization, Reinforcement learning, Battery life, Flight time.

1. INTRODUCTION

Technology is overpowering our lives now a day. Autonomous Vehicle is one of those leading technologies which are emerging. Unmanned aerial vehicle (UAV) is one of those autonomous vehicles which are useful and essential in current scenarios. UAVs are serving human life with different capabilities and applications in different sectors. UAVs can be useful for commercial, recreational and public uses. However, 90 percent of UAV production is meant for military purposes. But recently non-military drone market is expected to increase its sale in coming decades [1-2]. Service providers and delivery industries are looking for feasible yet smart delivery options [3]. UAVs can also be used for different monitoring purposes e.g. water sampling [4], land slide [5] and any volcanic eruptions [6-7] and civil health [8].

The areas where human cannot reach can possibly accessed by UAV for communication and radio access [9-11]. But, there are some technical challenges which restrict the expansion of commercial UAV utilization. Batter constraint is one of the major challenges which are required to address for successful implementation of UAV. On other hand, duration of flight of UAV is an important consideration and it is very much associated with battery life. So, the easy and basic solution is to increase in battery size or change in battery materials. But both the solutions are appeared to be unfeasible as the increase in battery size turns into increase in weight, which reduces the capability of flight time and it may also turn into cost increment of the device. The alternative way is to optimize the battery usage thus sustainable battery life can be ensured. To optimize the battery life, the drone needs to find take decision about the optimum route to reach at the destination. So, optimum route
selection becomes the key thing in this regard. In practical, it is very tough to find out the optimal path for routing the UAV as different obstacles, winds become the major issue. It becomes inappropriate to apply analytical techniques and classical law of guidance. In generic solution, dynamic program and any search based techniques are suggested for global path planning. But, there chance that it may suffer from Bellman’s curse of dimensionality [12]. In the context of flight time of UAV, figure 1. is drawn to understand the UAV during its flight.

It is studied that reinforcement learning [13] is different from supervised learning. In supervised learning, supervisor provides the desired output and the supervisor is well-known about the system model. In contrast, no supervision is there for RL and it depends upon the environment at the time of learning process though there is no prior information given to the system [13]. RL provides conventional search based methods, which involves discrete system space and control input space. A suitable framework is provided which includes vehicles motion model environmental disturbances. To achieve better path optimality under known disturbance condition path planning algorithm with optimal computational cost could be very much significant. In this paper, a reinforcement learning based path planning algorithm is proposed which may help to understand the feasible path ensuring optimal computational cost for UAV.

![Fig 1. An Unmanned Aerial Vehicle during its flight time](image)

2. SIGNIFICANT RELATED WORK

Recent days, an extensive work is going on path planning and it is becoming a primary research topic in vehicles guidance and robotics. The potential field method [14] is one of the popular approaches to provide guidance to vehicles. The vehicle’s routing path is computed after considering the artificial potential fields which represents the obstacles and destination. Graph search techniques are applied for guiding autonomous under water vehicles [15]. A path is selected by using the graph, where collisions do not take place. Genetic algorithm is useful to guide autonomous vehicles [16]. Real time disturbances are considered in path panning and dynamic computation approach is used to find optimized path to save the battery life [17-18].

The typical grid based search algorithm such as A* and Dijkstra’s are also reported with modified version [19-20]. Wind is also equally important consideration for UAV’s [21-23]. In case of autonomous vehicles, different resistance are considered in path planning and using conventional or modified versions of generic algorithms are applied to generate paths. Some of the famous algorithm e.g. A*, rapidly exploring random tree (RRT) are used to find out the path for autonomous vehicles.
Fig 2. Flow diagram of reinforcement learning

RL algorithms are suitable as trial-error feedbacks are used for achieving optimized solution. From various review, it is studied that these algorithms are providing feasible framework for the problem which was thought to be a great challenge [13]. But still there is a need of improving the algorithm to ensure more optimized path thus increased battery life and flight time.

3. REINFORCEMENT LEARNING (RL) IN PATH PLANNING:

There are some basic components of the RL algorithm, which includes agent, an environment, states, action keywords, policies and action value function. The figure 2 describes the block diagram of reinforcement learning. The environment basically defines the situation which gives reward to agent. States can be defined as parameters which describe the agent and the environment. The state of the agent is changed by action. Rewards are evaluated depending upon the environment after determining the pros and cons of the applied action. At different states actions are selected based on the policies.

RL can be proposed to find optimal solution. When an agent chooses desired action, a reward is given, this process repeats. This encourages the agent to maximize its rewards. Based on the temporal difference learning [13] L can be employed as the learning algorithm in this following study. In L learning, \( L(p, f) \) is the action values function. \( L(p, f) \) is expected reward from present state to final state. Where \( p \) represents present state and \( f \) denotes the final state.

The RL problem is described as searching for optimal solution, criteria for searching optimal solution is the selection of good action in every step of the learning. The learning rate and process may differ based on how \( L(p, f) \) is calculated to select an action. The value \( L(p, f) \) can be updated in this manner.

\[
L(p, f) \leftarrow (1-\alpha)L(p, f) + \alpha(r + \gamma \max L(p', f'))
\]  

Where, \( r \) denotes rewards after the \( p, f \) and \( p', f' \) are the next state and action. Learning rate is denoted as \( \alpha \) and \( \gamma \) represents discontinuing factor which is used to use weight the learned result near the current time step. The \( \alpha \) is always positive constant and here the value is 0.1 and \( \gamma \) is set to 1.

Using the equation 1, the value of \( L \) can be updated. Since, \( L \) converges to the optimal solution when the maximum \( L \) value is obtained the next step. By applying temporal difference method, next \( L \) value can be computed from the current \( L \) value. To select the maximum \( L \) which leads to optimal action is known as greedy action. However, agent must explore other actions during different learning iterations. If agent only chooses greedy action, other actions cannot be explored. It does turn into optimal \( L \) value in the initial iterations. This is known as off policy method as different policies are considered when \( L \) is computed and action is selected. To compute current \( L \), the next maximum \( L \).
value is used. By keeping the balance between exploring next state and exploiting current states, $L$ can be computed effectively. In this investigation, $\varepsilon$ greedy policy is used for optimal action depending on the probability of $\alpha$ \cite{24}.

4. PATH/ROUTE PLANNING FORMULATION

The kinetic vehicle model describes the range of vehicle motion. In route planning, it is considered during the discretization of the motion states and control action in continuous space by applying RL techniques. A tile coding process is used for incorporating continuous state space variable \cite{13}. It is clearly understood from the name of the technique that the receiver fields are clubbed into section of the state space. Using a set of square tiles, the position state variables are discretized into square domains. The vehicles direction state is discretized into finite interims. Importantly, the direction of the vehicle is maintained in a position tile where the vehicle is located at that moment. When the vehicle reaches the boundary of the adjacent tile, a new action is determined. $L(p, f)$ is also updated along with the position and direction after selecting the new state action \cite{13}. For controlling the yaw rate of the vehicle an individual policy is in the action. The amount of time the agent remains in each tile is considered as a “penalty”. For each successful journey, the sum of penalties considered as elapsed time for the journey from initial point to terminal point. The total reward can be determined by multiplying some of penalties with -1. If the total reward is increased, the total optimal time path is achieved. $L$ is constituted with different actions which will determine the optimum time path between starting and destination point. In $L$ learning generally it is assumed that the initial $L$ value is “zero”. It implies that no information about $L$ is provided. If any prior information is available, the $L$ function may be initialized as non-zero function. The nonzero value of $L$ function can affect the performance of $L$ learning; basically the convergence of the function will be disturbed.

5. PATH/ROUTE SMOOTHING

The degree of smoothness in the determined path and angle of trajectories depend upon the discretized resolution. There is relation between computational efficiency and smoothness of the path. The geometric route cannot be smoothening directly as motion constraints and wind current effects are ignored during these process. The vehicle speed profile is relative to the place as each tile is affected by different parameters of the actions. In this study, yaw angles are smoothed rather that smoothening the path itself. The end point is shifted to predefined destination by using some iterative schemes. After smoothening the yaw angles, they are stretched or compressed slightly along the time and angle axes. The Correction factors $T_\phi$ and $\theta_c$ are considered as variations on time and angle axes respectively. By using the steepest descent method \cite{25}, the appropriate value of $T_\phi$ and $\theta_c$ can be calculated.

6. PROPOSED PATH/ROUTE PLANNING ALGORITHM FOR UAV ROUTING APPLICATION

The algorithm proposed for the UAV to ensure optimized path selection, higher battery life and flight time is discussed.

Define $L(p, f)$ with a value, repeat for each iteration

```
Initialize $p$ in each repeat for each step in each iteration

Choose $f$ from $p$ using the policy derived from $L$

Take action $p$, observe $r$, $f'$

$$L(p, f) \leftarrow L(p, f) + \alpha(r + \gamma) \max_{f'} L(p', f') - L(p, f)$$

$$S \leftarrow S'$$
```

4
Until $f$ is the terminal.

**Until convergence**

**Smooth** path by applying average

Initialize $T_c, \theta_c$

While $e_{\varepsilon} > \varepsilon$

Define $e_{\varepsilon}$ distance is predefined and shifted end point

$$T_c \leftarrow T_c - \alpha \frac{\partial f(T_{\varepsilon}, \theta_{\varepsilon})}{\partial T_{\varepsilon}}$$

$$\theta_c \leftarrow \theta_c - \beta \frac{\partial f(T_{\varepsilon}, \theta_{\varepsilon})}{\partial \theta_{\varepsilon}}$$

End while

End post-processing

7. **CONCLUSION:**

The proposed work is focused on a sustainable method for path or route optimization of an unmanned aerial vehicle and the novelty of the work lies on the introduction of reinforcement based algorithm for UAV routing application. The disturbances like wind current and obstacles are considered to provide enhanced routing efficiency in real-time application. Herein, a suitable reinforcement learning based algorithm is proposed for optimizing the path which in turns may increase the battery life and ensure higher flight time. Based on the proposed algorithm for the particular UAV applications, the simulation can be done and a comparative study can be made with respect to other existing algorithms for UAV application.

**REFERENCES**

[1] Global Industry Analysts, Inc. 2016 *Commercial Drones Market Trends*. Accessed: Jan.4,2018. Available:http://www.strategyr.com/MarketResearch/Commercial_Drones_Market_Trends.asp

[2] Price water house Coopers. 2016 *Global Market for Commercial Applications of Drone Technology* Valued at Over $127 BN. Accessed: Jan. 4, 2018. [Online]. Available: http://press.pwc.com/News-releases/global-market-for-commercial-applications-of-drone-technology-valuedat-over–127-bn/s/ac04349e-c40d-4767-9f92-a4d219860ed2

[3] M. Gharibi, R. Boutaba, and S. L. Waslander, 2016 *Internet of drones*, IEEE Access vol. 4, pp. 1148–1162.

[4] J.-P. Ore, S. Elbaum, A. Burgin, and C. Detweiler, Dec. 2015 Autonomous aerial water sampling, *J. Field Robot.*, vol. 32, no. 8, pp. 1095–1113.

[5] U. Niethammer, M. R. James, S. Rothmund, J. Travelletti, and M. Joswig, Mar. 2012 UAV-based remote sensing of the Super-Sauze landslide: Evaluation and results *Eng. Geol.*, vol. 128, pp. 2–11.

[6] G. Astuti, G. Giudice, D. Longo, C. D. Melita, G. Muscato, and A. Orlando, 2009 An overview of the ‘volcan project’: An UAS for exploration of volcanic environments, *J. Intell. Robot. Syst. Theory Appl.*, vol. 54, nos. 1–3, pp. 471–494.

[7] G. Zhou and D. Zang, Jul. 2007 Civil UAV system for earth observation, in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, pp. 5319–5322.

[8] J. Johnson, E. Basha, and C. Detweiler, Apr. 2013, Charge selection algorithms for maximizing sensor network life with UAV-based limited wireless recharging, in *Proc. IEEE 8th Int. Conf. Intell. Sensors, Sens. Netw. Inf. Process.(ISSNIP)*, pp. 159–164.

[9] W. Shi et al., 2018 Multiple drone-cell deployment analyses and optimization in drone assisted radio access networks, *IEEE Access*, vol. 6, pp. 12518–12529.

[10] E. Lee, C. Choi, and P. Kim, 2017 Intelligent handover scheme for drone using fuzzy inference
systems, IEEE Access, vol. 5, pp. 13712–13719.

[11] H. C. Nguyen, R. Amorim, J. Wigard, I. Z. Kovács, T. B. Sørensen, and P. E. Mogensen, 2018, How to ensure reliable connectivity for aerial vehicles over cellular networks, IEEE Access, vol. 6, pp. 12304–12317.

[12] Bryson B 1975 Applied optimal control: optimization, estimation and control (CRC Press)

[13] Sutton RS, Barto AG, 1998, Introduction to Reinforcement Learning (MIT Press).

[14] Carroll KP, McClaran SR, Nelson EL, Barnett DM, Friesen DK, Williams GN June 02–03, 1992 AUV path planning: an A approach to path planning with consideration of variable vehicle speeds and multiple, overlapping, time-dependent exclusion zones. In: Proceedings of the Symposium on AUV Technology, Washington, DC, pp 79–84.

[15] Warren CW, 1990, A technique for autonomous underwater vehicle route planning. IEEE J Ocean Eng 15(3):199–204.

[16] Sugihara K, Yuh J, 1997, GA-based motion planning for underwater robotic vehicle. In: Proceedings of the 10th International Symposium on Unmanned Untethered Submersible Technology, Durham, NH, USA, pp 406–415.

[17] Alvarez A, Caiti A, 2002, Interactions of autonomous underwater vehicles with variable scale ocean structures. In: Proceedings of the IFAC World Conference Systems, Barcelona, Spain.

[18] Alvarez A, Caiti A, Onken R, 2004, Evolutionary path planning for autonomous underwater vehicles in a variable ocean. IEEE J Ocean Eng 29(2):418–429.

[19] Garau B, Alvarez A, Oliver G, April 18–22, 2005, Path planning of autonomous underwater vehicles in current fields with complex spatial variability: an A* approach. In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp 194–198.

[20] Lee T, Chung H, Myung H, June 06–09, 2011, Multi-resolution path planning for marine surface vehicle considering environmental effects. In: 2011 IEEE OCEANS, Santander, Spain, pp 1–9.

[21] Shao W, Zhou P, Thong SK 2012 Development of a novel forward dynamic programming method for weather routing. J Mar Sci Technol 17(2):239–251.

[22] Lin YH, Fang MC, Yeung RW 2013 The optimization of ship weather-routing algorithm based the composite influence of multi-dynamic elements. Appl Ocean Res 43:184–194.

[23] Padhy CP, Sen D, Bhaskaran PK 2008 Application of wave model for weather routing of ships in the North Indian Ocean. Nat Hazards 44(3):373–385.

[24] Yoo, B., & Kim, J. 2016 Path optimization for marine vehicles in ocean currents using reinforcement learning. Journal of Marine Science and Technology, 21(2), 334-343.

[25] Chapra SC 2010 Numerical Methods for Engineers, 6th edn. (McGraw-Hill).