Reinforcement Learning-Based Upset Recovery Policy for Aircraft Achieved in Two Stages

Huanhui Cao, Weifeng Zeng, Hantao Jiang, Hao Hu, Chaoran Li, Wenjie Lu, and Hao Xiong*
School of Mechanical Engineering and Automation
Harbin Institute of Technology
Shenzhen, China
xionghao@hit.edu.cn

***Abstract*—Aircraft upset situations are the highest risk to civil aviation. Thus, a reliable upset recovery policy is necessary for aircraft. In this paper, a strategy to achieve a reinforcement learning-based upset recovery policy that takes time of recovery and loss of altitude into account in two stages is proposed for aircraft to recover from an arbitration upset situation to level flight. According to the proposed strategy, the training of an RL-based upset recovery policy for a certain aircraft is supposed to speed up. Based on the idea of obtaining an RL-based upset recovery policy in two stages, an algorithm to achieve a Twin Delayed Deep Deterministic Policy Gradient (TD3)-based upset recovery policy in two stages is developed. Experiments of stall recovery and spin recovery are conducted based on X-Plane 11 to validate the effectiveness of the proposed strategy and the performance of the achieved upset recovery policy.**

***Keywords*—aircraft upset recovery, reinforcement learning, twin delayed deep deterministic policy gradient, two stages**

I. Introduction

Aircraft upset incidents are the highest risk to civil aviation from decades ago to now [1]. To this end, several researchers have devoted to address the aircraft upset issue [2]–[4]. Yildiz et al. [2] proposed a finite-state conditional switching structure to achieve the recovery from loss-of-control conditions. To deal with stall recovery, Cunis et al. [3] developed a Loss of Altitude (LOA) minimizing approach based on economic model predictive control. By solving a trajectory optimization problem via direct multiple shooting method, aircraft spin recovery was addressed in [4]. Although the above-mentioned approaches can address nonlinear dynamics of aircraft, these approaches may fail to handle the high complexity of upset situations [5].

Reinforcement Learning (RL) is an approach that can deal with high complexity problems without an explicit model of the problem [6], [7]. Thus, RL is appropriate to be applied to develop upset recovery policy for aircraft suffering complex upset situation [5]. Dutoi et al. combined robust control and RL to address spin recovery problem [8]. The spin recovery performance of the achieved policy outperforms the policy obtained based on skilled pilots in some cases. Nonetheless, this study compressed the action space to ease the computation burden, leading to a limitation on agile spin recovery. Kim et al. [5] developed an RL-based recovery policy including nine actions for unusual attitude recovery and 27 actions for angular rate arrest for stable flat spin. By comparing to an optimal solution, the performance of the proposed RL-based policy was verified. However, the training of the RL-based policy was not detailed in [5]. Zhu et al. [9] applied deep deterministic policy gradients (DDPG) and deep Q-network (DQN) to achieve spin recovery policies and propose an exploring mechanism that dynamically selects between deterministic and stochastic exploration for DDPG. Whereas this study did not consider the uncertainties of the aircraft and the environment.

This paper proposes to achieve upset recovery policy for aircraft in two stages. The major contributions of this paper are as follows.

- A strategy to achieve RL-based upset recovery policy in two stages, which includes a pre-train stage and a fine-tuning stage, is proposed for aircraft. The achieved upset recovery policy address different upset situations of an aircraft, rather than stall or spin only. According to the proposed strategy, the training of a RL-based upset recovery policy for a certain aircraft can be speed up.
- Experiments are conducted based on a Cessna172SP aircraft in X-Plane 11 to validate the effectiveness of the proposed strategy and the performance of the achieved upset recovery policies.

The rest of this paper is organized as follows. Section II demonstrates the preliminaries, including typical upset situations of aircraft, loss of altitude, reinforcement learning, and Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm. In section III, a strategy to achieve RL-based upset recovery policy in two stages is proposed. Section IV presents experiments to illustrate the proposed strategy and validate the effectiveness of the proposed strategy. Finally, section V summarizes this paper.

II. Preliminaries

A. Upset Situations of Aircraft

An upset situation of aircraft refers to an abnormal mode of the nonlinear dynamics that shows significantly altered steady-state responses [3], such as stall and spin.

Stall [10]. In aerodynamics and aviation, stall is a condition such that lift begins to decrease suddenly if the angle of attack

*Corresponding author.

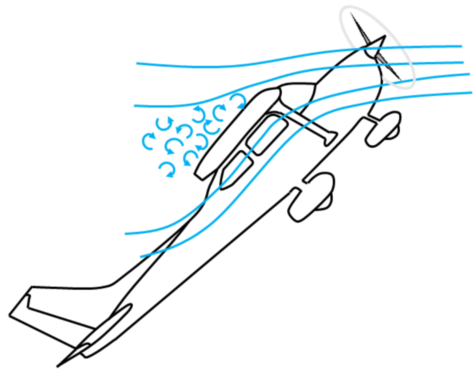

Fig. 1. Stall of an aircraft

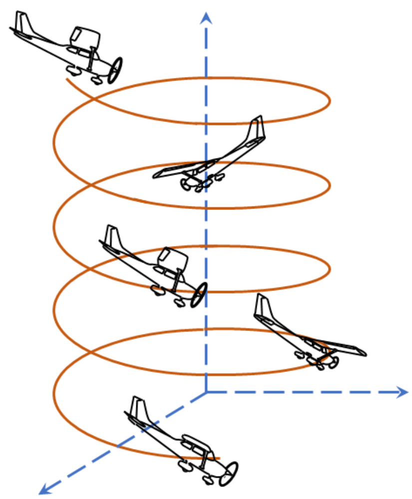

Fig. 2. Spin of an aircraft

of aircraft increases beyond a certain point. Stall of an aircraft is shown in Fig. 1.

Spin [11]. Spin is a nonlinear post-stall phenomena in which an aircraft develops a high rotational-rate and descends almost vertically in a helical trajectory, as shown in Fig. 2. For an aircraft in spin, one wing is stalled more than the other. The more stalled wing is with less lift and more drag than the other, leading to the auto-rotation and subsequent rapid decent of the aircraft.

B. Loss of Altitude

LOA [3] is a significant performance index for an upset recovery policy. LOA can be exploited to enlarge the operational envelope of aircraft, particularly at low altitudes. The LOA has been used in several previous researches [12], [13] to evaluate control policies of aircraft. In this study, the LOA is an significant index to measure the performance of achieve upset recovery policy.

C. Reinforcement Learning and Twin Delayed Deep Deterministic Policy Gradient Algorithm

Reinforcement learning studies the paradigm of an agent interacting with the environment aiming to learn policies that maximize accumulated rewards. At time step t, the agent selects an action $a \in \mathcal{A}$ based on the current state $s \in \mathcal{S}$ with respect to its policy $\pi : \mathcal{S} \mapsto \mathcal{A}$. The agent receives a reward r and the state transfers to a new state s'. The agent aims to maximize the accumulated rewards $R_t = \sum_{i=t}^{T} \gamma^{i-t} r(s_i, a_i)$, where γ is a discount factor.

TD3 algorithm [14] is a RL algorithm proposed for agents with continuous states and actions based on an actor-critic architecture. TD3 applies three approaches to address the overestimation issue of the critic network.

Smooth target policy. In order to reduce the variance caused by over-fitting, TD3 adds a clipped random noise to the value of target critic network. After adding the clipped noise, target action network is clipped to lie in feasible action range.

Apply a pair of critic networks. TD3 includes two critic networks. In the Bellman error loss functions, one should take the smaller value of these two target critic networks. In this way, the overestimation of the critic network is alleviated.

Delay actor updates. TD3 updates the actor network less frequently than the critic network. The actor network is updated after a certain number of time steps, while the critic network is update at every time step.

III. Achieve Reinforcement Learning-Based Upset Recovery Policy in Two Stages

In this section, a strategy to achieve RL-based upset recovery policy in two stages is proposed for aircraft.

A. Problem Formulation

In aerodynamics and aviation, upset situations can refer to a variety of abnormal situations. An upset recovery policy is expected to recover aircraft from an upset situation with minimum LOA within the shortest possible time in practice. Assume that an aircraft can adjust its state by aileron, elevator, rudder, and throttle, as shown in Fig. 3. According to studies of multi-objective reinforcement learning [15], The problem addressed by an upset recovery policy can be expressed as

$$\begin{aligned} &\min \ \{w_{LOA}||LOA(a_i)|| + \sum_{m=1}^{M} \omega_m ||s_m(a_i) - s_m^*||\} \\ &s.t. \ \ a_i^L \le a_i \le a_i^H, \quad i = 1,2,3,4 \end{aligned} \tag{1}$$

where $a_i (i = 1, 2, 3, 4)$ denotes the action of elevator, aileron, rudder, throttle of an aircraft, respectively. a_i^L and a_i^H are the lower bound and the upper bound of a_i, respectively. $s_m(a_i)(m = 1, 2, ..., M)$ denotes the state of the aircraft. s_m^* denotes a target state. $LOA(a_i)$ denotes LOA function. $w_m(m = 0, 1, ..., M)$ are the weight coefficients of the state $s_m(a_i)$.

For an aircraft, an upset recovery policy is not supposed to control the aircraft to certain locations usually, so latitude

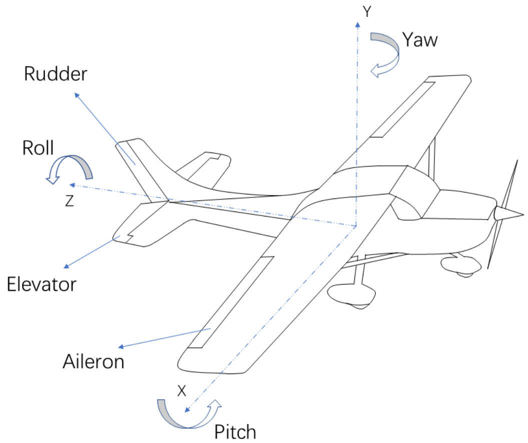

Fig. 3. Control surfaces of an aircraft

and longitude can be ignored in the study of upset recovery policies. One can define the state space of an upset recovery policy as $\mathcal{S} = \{\omega, \kappa, \xi, p, q, r, h, v\}$ and has $s_m(a_i) \in \mathcal{S}$. ω, κ, ξ, p, q, r, h, v denote the pitch, roll, yaw, pitch angular velocity, roll angular velocity, yaw angular velocity, altitude, and airspeed of the aircraft, respectively.

B. Pre-train Stage and Fine-tuning Stage of RL-Based Upset Recovery Policy

The dynamics of an aircraft is with complexity and uncertainty caused by aerodynamics, weather, fuel, and payload. Upset recovery policies based on robust control and the knowledge of dynamics of the aircraft cannot achieve the optimal performance [5]. RL is a promising upset recovery scheme in the sense of complexity and uncertainty addressing. In view of the uncertainties of upset recovery for aircraft such as aerodynamics, weather, fuel, and payload, a strategy to achieve an upset recovery policy in two stages (i.e., pre-train stage and fine-tuning stage) is proposed, as shown in Fig. 4.

In the pre-train stage, a general upset recovery policy is learned based on RL for an aircraft with parameters (e.g., aerodynamics, weather, fuel, and payload) disturbed in certain ranges randomly. The general upset recovery policy is sub-optimal but is with the adaptability to aircraft with parameter uncertainties. In the fine-tuning stage, a specific upset recovery policy is learned via RL by the current aircraft whose aerodynamics, fuel, and payload are determined, based on the general upset recovery policy. Based on the general upset recovery policy achieved in the pre-train stage, rather a random initial policy, the training of the specific upset recovery policy for the current aircraft can speed up.

C. Reward Function

The design of reward function is crucial for RL since it affects not only the convergence but also the quality of the convergent point of the RL. Inspired by the idea of rewarding

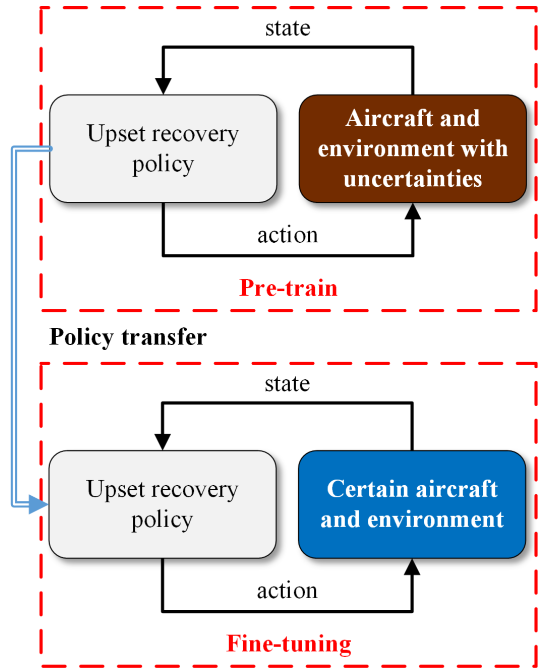

Fig. 4. Flowchart of the two stages of RL-based upset recovery policy.

shaping [16], we define the reward function of RL-based upset recovery policy as follows:

$$R(s, a, s') = T(s') + P(s) + C(s, a, s') \tag{2}$$

where $T(s')$ is a termination reward. $P(s)$ is a state-related reward. $C(s, a, s')$ is an attitude-related reward.

The termination reward is the reward obtained when an upset recovery policy is success to recover an aircraft or makes the aircraft crushed eventually. The termination reward is defined as:

$$T(s') = \begin{cases} T_1, & \text{if the aircraft is crushed} \\ T_2, & \text{if success to recover the aircraft} \\ 0, & else \end{cases} \tag{3}$$

where T_1 is a negative reward while T_2 is a positive reward.

The state-related reward $P(s)$ is expressed as:

$$P(s) = w_{LOA}||LOA^n|| + (\sum_{m=1}^{8} w_m||s_m^n - s_m^{*,n}||) \tag{4}$$

where s_m^n, $s_m^{*,n}$, LOA^n are the normalization of state s_m, target state s_m^*, LOA of an aircraft, respectively. $w_m(m = 1, ..., 8)$ is a non-positive weight coefficient of the difference between s_m^n and $s_m^{*,n}$. w_{LOA} is a non-positive weight coefficient of LOA.

$C(s, a, s')$ depends on the attitude and angular velocity of an aircraft. If the angular velocity tends to decrease the difference between the attitude of the aircraft and the target attitude, a positive reward c_1 will be received. Otherwise, a negative reward c_2 will be received.

D. TD3-Based Upset Recovery Policy Achieved in Two Stages

Based on the strategy to achieve RL-based upset recovery policy in two stages presented in section III-B and the reward function presented in section III-C, we can develop a strategy to achieve TD3-based upset recovery policy by applying TD3 algorithm in both the pre-train stage and in the fine-tuning stage. The pseudo-code of the TD3 algorithm applied to achieve upset recovery policy is shown in Algorithm 1.

Although the TD3 algorithm is used in both the pre-train stage and in the fine-tuning stage, applications of the TD3 algorithm are different from two aspects - initialization and parameter uncertainty. For the initialization aspect, the upset recovery policy is initialized randomly in the pre-train stage while, in the fine-tuning stage, the upset recovery policy is initialized to the general upset recovery policy achieved in the pre-train stage. For the parameter uncertainty aspect, the environment is with parameters distributed randomly in certain ranges in the pre-train stage, but the environment is with parameters determined by the current aircraft.

IV. Experiments

A. Environment Setup

To evaluate the proposed strategy to achieve upset recovery policy in two stages, we conduct experiments based on X-Plane 11, a flight simulation environment. X-Plane 11 is a professional flight simulation environment, involving various types of aircraft ranging from general aircraft to commercial aircraft. X-Plane has been used by several researchers [17]–[19] to study the design, control, and guidance of aircraft. A Cessna172SP is included in experiments to study the performance of the upset recovery policy achieved in two stages, as shown in Fig. 5. Moreover, the wind condition in the experiments is adjusted to achieve a general upset recovery policy in pre-train stage.

X-Plane 11 is integrated to TensorFlow-gpu 2.4 in this study. A personal computer with 64-bit Windows 10 operation system, 32 gigabytes memory, a 3.2GHz Corei7 CPU, and a RTX3070 GPU is used to validate the proposed strategy and the performance of achieved upset recovery policies.

Fig. 5. Aircraft included in experiments - Cessna172SP

Algorithm 1: Algorithm to achieve a TD3-based upset recovery policy for aircraft

Input: Empty replay buffer $\mathcal{D}$

if *pre-train stage* **then**
- Initialize actor network θ, critic networks ϕ_1 and ϕ_2;
- Initialize target networks: $\theta_{targ} \leftarrow \theta$, $\phi_{targ,1} \leftarrow \phi_1$, $\phi_{targ,2} \leftarrow \phi_2$;
- Set parameters of aircraft and environment randomly in certain ranges;

if *fine-tuning stage* **then**
- Load actor network θ, critic networks ϕ_1 and ϕ_2, and target networks θ_{targ}, $\phi_{targ,1}$, $\phi_{targ,2}$ achieved in pre-train stage;
- Set parameters according to current aircraft and environment;

for *episode=1 to M* **do**
- Receive initial observation state s_{ini};
- Normalize the initial state s_{ini};
- **for** *t=1 to T* **do**
  - Select action with exploration noise $a \sim \pi_\theta(s) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \delta)$;
  - Convert action a into the aircraft maneuver a_m;
  - Execute a_m, and obtain reward r and new state s';
  - Normalize the state s';
  - Store transition (s, a, r, s', d) in $\mathcal{D}$;
  - Sample a mini-batch of transitions $\mathcal{B} = \{(s, a, r, s', d)\}$ from $\mathcal{D}$;
  - Compute action $a'(s') = clip(\pi_{\theta_{targ}}(s') + clip(\epsilon, -c, c), a_{Low}, a_{High})$;
  - Compute the value of target critic network $y(r, s', d) = r + \gamma(1-d)\min_{i=1,2} Q_{\phi_{i,targ}}(s', a')$;
  - Update critic network by one step of gradient ascent using: $\nabla_{\phi_i} \frac{1}{\mathcal{B}} \sum_{(s,a,s',r,d)\in\mathcal{B}} (Q_{\phi_i}(s,a) - y(r,s',d))^2$;
  - **if** *t mod p_{delay}=0* **then**
    - Update actor network by one step of gradient ascent using: $\nabla_\theta \frac{1}{\mathcal{B}} \sum_{s\in\mathcal{B}} Q_{\phi_1}(s, \pi_\theta(s))$;
    - Update target networks: $\phi_{targ,i} \leftarrow \tau\phi_{targ,i} + (1-\tau)\phi_i \quad i = 1, 2$; $\theta_{targ} \leftarrow \tau\theta_{targ} + (1-\tau)\theta$

B. Training Setup

With the above-mentioned environment set up, we set the TD3 algorithm included in the TD3-based upset recovery policy as follows. The learning rate is 1.0×10^{-4}. The discount factor γ is set to 0.99. The actor network has three hidden layers that have 64, 64, and 32 units, respectively. The critic network has two hidden layers with 64, 64 units, respectively. The time step T is defined as 500 and the minibatch size $\mathcal{B}$ is 64. The number of episodes M is 800 in pre-train stage and is 50 in fine-tuning stage.

The feasible range of the states of the aircraft included in experiments are listed in Table I. The parameters of the reward function used in experiments are presented in Table II.

Table I. The feasible range of the states of aircraft

variable	value	variable	value
ω,κ	$[-180°, 180°]$	ξ	$[0°, 360°]$
p, q, r	$[-90°/s, 90°/s]$	h	$[0\ m, 5000\ m]$
v	$[0\ m/s, 120\ m/s]$		

Table II. The parameters of the reward function

variable	value	variable	value
T_1	-1000	T_2	500
p_0	-1.0	p_1, p_2, p_3, p_6	-0.1
p_4, p_5, p_7, p_8	-0.5	c_1	1.0
c_2	-1.0		

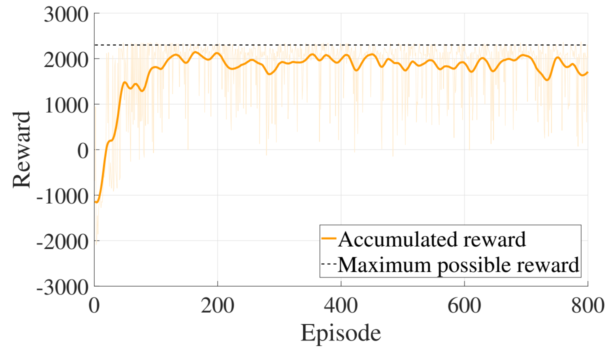

Fig. 6. Reward achieved in the pre-train stage

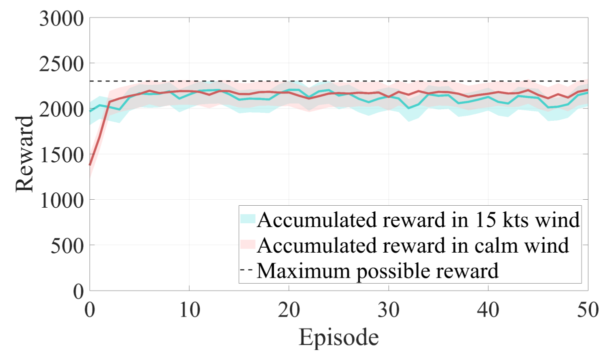

Fig. 7. Rewards achieved in the fine-tuning stage

C. Training of Upset Recovery Policies

Based on the environment set up and training set up presented above, we conduct the training of upset recovery policies according to the proposed strategy. In pre-train stage, it is assumed that the fuel and payload of aircraft and the weather are not determined. Thus, the weight of the aircraft distributes randomly from 781 kilograms to 1160 kilograms and wind speed distributes randomly from 0 knots to 100 knots. In fine-tuning stage, it is assumed that the weight of aircraft and the weather are determined. The initial upset recovery policy is the policy achieved in the pre-train stage. The accumulated rewards achieved in the pre-train stage and in the fine-tuning stage are shown in Fig. 6 and Fig. 7.

As shown in Fig. 6, in pre-train stage, upset recovery policy converges in 180 episodes. The accumulated reward reaches approximately 2000 on average after convergence. It is shown that the upset recovery policy achieved in the pre-train stage is sub-optimal with respective to the maximum possible reward. Besides, one can see that the upset recovery policy performs poorly in some episodes after the convergence.

In fine-tuning stage, the weight of aircraft and the weather are determined. We train upset recovery policy based on the policy achieved in pre-train stage, both in 15 knots wind and in calm wind. As shown in Fig. 7, upset recovery policy converges in four episodes both in wind and in calm wind. This shows that train of upset recovery policy can be speed up if an upset recovery policy achieved in pre-train stage is available. Also, one can see that upset recovery policy can approach the maximum possible reward both in wind and in calm wind with fine-tuning.

D. Evaluation of Upset Recovery Policies

The performance of upset recovery policies achieved according to the proposed strategy is evaluated based on stall recovery and spin recovery. Proportional-Integral-Differential (PID) controller that has been applied on the upset recovery of aircraft [20], [21] is well tuned and included in the evaluation working as references.

Stall recovery. TD3-based upset recovery policy and PID-based upset recovery policy are used to address stall recovery of an aircraft in wind and in calm wind. The roll, pitch, airspeed, and altitude of the aircraft are shown in Fig. 8 - Fig. 10. Both the TD3-based upset recovery policy and the PID-based upset recovery policy can recover the aircraft from stall to level flight both in wind and in calm wind. However, according to Fig. 11, the TD3-based upset recovery policy can recover the aircraft with a smaller LOA than the PID-based upset recovery policy can both in wind and in calm wind.

Spin recovery. TD3-based upset recovery policy and PID-based upset recovery policy are used to deal with spin recovery of an aircraft in wind and in calm wind. The roll, pitch, airspeed, and altitude of the aircraft are shown in Fig. 12 - Fig. 14. Both the TD3-based upset recovery policy and the

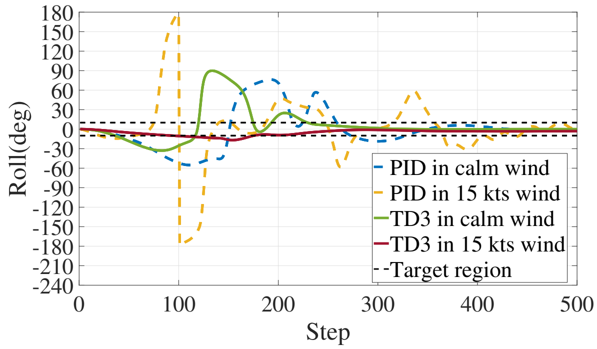

Fig. 8. The roll of the aircraft in stall recovery

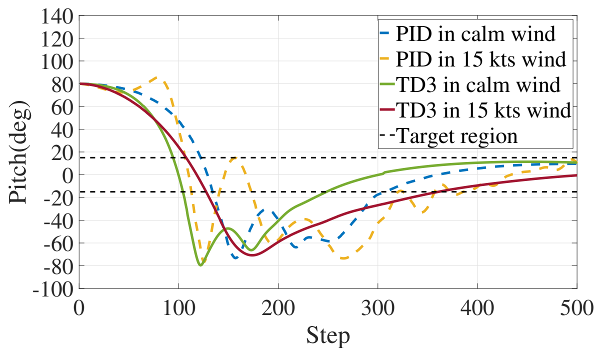

Fig. 9. The pitch of the aircraft in stall recovery

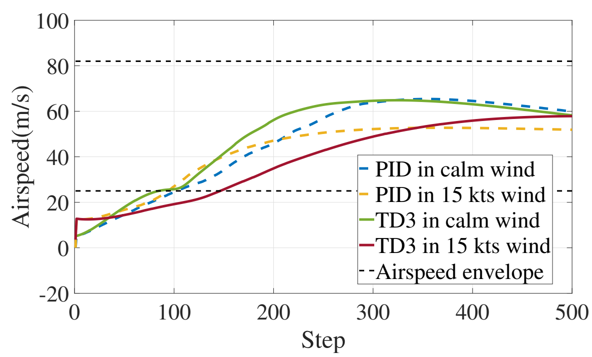

Fig. 10. The airspeed of the aircraft in stall recovery

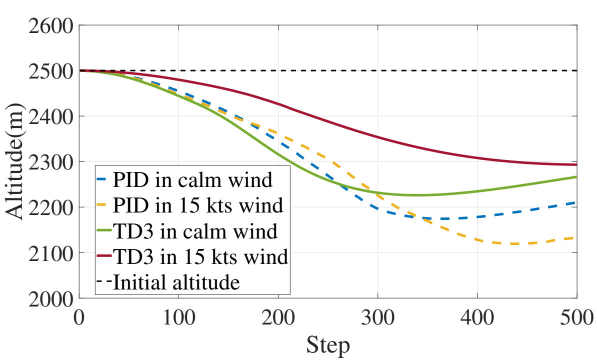

Fig. 11. The altitude of the aircraft in stall recovery

PID-based upset recovery policy can recover the aircraft from spin to level flight both in wind and in calm wind. As shown in to Fig. 15, the TD3-based upset recovery policy can recover the aircraft with a smaller LOA than the PID-based upset recovery policy can in 15 kts wind.

According to the results of the experiments of stall recovery and spin recovery, one can see that TD3-based upset recovery policy can recover an aircraft into its flight envelope. The effectiveness of the RL-based upset recovery policy is validated. Moreover, TD3-based upset recovery policy achieves better performances in LOA and time of recovery than well-tuned PID-based upset recovery policy.

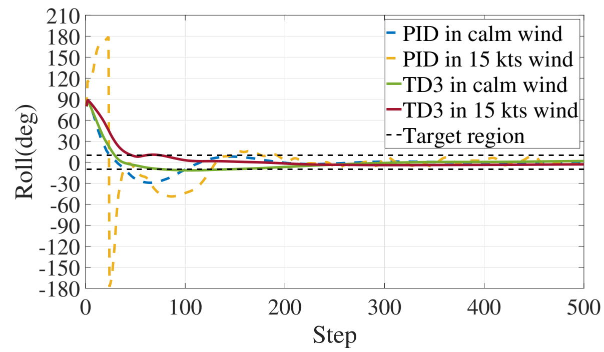

Fig. 12. The roll of the aircraft in spin recovery

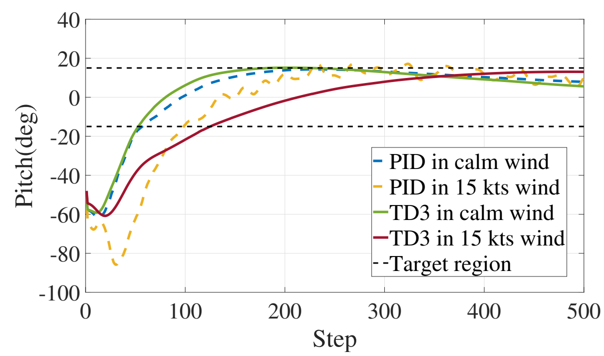

Fig. 13. The pitch of the aircraft in spin recovery

V. Conclusion

In this paper, a strategy to achieve a RL-based upset recovery policy in two stages was proposed to recover an aircraft from an arbitrary upset situation to level flight taking time of recovery and LOA into account. According to the proposed strategy, an algorithm to achieve a TD3-based upset recovery policy was developed. To demonstrate the implementation and verify the effectiveness of the proposed strategy, training of upset recovery policies and experiments of stall recovery and spin recovery were conducted based on a Cessna172SP in X-Plane 11. The results of training and experiments validate the effectiveness of the proposed strategy and the performance of achieved upset recovery policy.

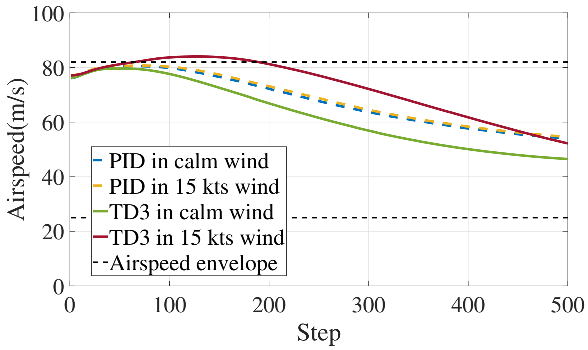

Fig. 14. The airspeed of the aircraft in spin recovery

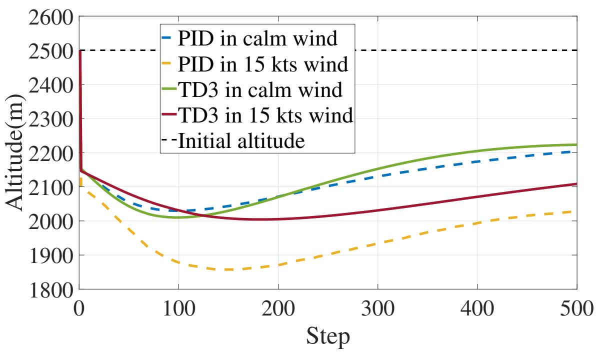

Fig. 15. The altitude of the aircraft in spin recovery

REFERENCES

[1] Christine M Belcastro, John V Foster, Gautam H Shah, Irene M Gregory, David E Cox, Dennis A Crider, Loren Groff, Richard L Newman, and David H Klyde. Aircraft Loss of Control Problem Analysis and Research Toward a Holistic Solution. *Journal of Guidance, Control, and Dynamics*, 40(4):733–775, 4 2017.

[2] Anil Yildiz, M Ugur Akcal, Batuhan Hostas, and N Kemal Ure. Switching Control Architecture with Parametric Optimization for Aircraft Upset Recovery. *Journal of Guidance, Control, and Dynamics*, 42(9):2055–2068, 4 2019.

[3] T Cunis, D Liao-McPherson, J Condomines, L Burlion, and I Kolmanovsky. Economic Model-Predictive Control Strategies for Aircraft Deep-stall Recovery with Stability Guarantees. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 157–162, 2019.

[4] D.M.K.K. Venkateswara Rao and Tiauw H Go. Optimization of aircraft spin recovery maneuvers. *Aerospace Science and Technology*, 90:222–232, 2019.

[5] Donghae Kim, Gyeongtaek Oh, Yongjun Seo, and Youdan Kim. Reinforcement Learning-Based Optimal Flat Spin Recovery for Unmanned Aerial Vehicle. *Journal of Guidance, Control, and Dynamics*, 40(4):1076–1084, 6 2016.

[6] Zhuangdi Zhu, Kaixiang Lin, and Jiayu Zhou. Transfer learning in Deep Reinforcement Learning: A survey. *arXiv*, pages 1–22, 2020.

[7] Hao Xiong and Xiumin Diao. Safety Robustness of Reinforcement Learning Policies: A View from Robust Control. *Neurocomputing*, 422:12–21, 2021.

[8] Brian Dutoi, Nathan Richards, Neha Gandhi, David Ward, and John Leonard. Hybrid Robust Control and Reinforcement Learning for Optimal Upset Recovery. In *AIAA Guidance, Navigation and Control Conference and Exhibit*, Guidance, Navigation, and Control and Co-located Conferences. American Institute of Aeronautics and Astronautics, 8 2008.

[9] Y Zhu, H Liu, B Ren, H Duan, X She, and Z Wu. A Model-free Flat Spin Recovery Scheme for Miniature Fixed-wing Unmanned Aerial Vehicle. In *2019 IEEE International Conference on Unmanned Systems (ICUS)*, pages 623–630, 2019.

[10] Ashraf Omran and Ayman Kassem. Optimal task space control design of a Stewart manipulator for aircraft stall recovery. *Aerospace Science and Technology*, 15(5):353–365, 2011.

[11] Bilal Malik, Jehanzeb Masud, and Suhail Akhtar. A review and historical development of analytical techniques to predict aircraft spin and recovery characteristics. *Aircraft Engineering and Aerospace Technology*, 92(8):1195–1206, 1 2020.

[12] D W Sparks and D D Moerder. Optimal aircraft control upset recovery with and without component failures. In *Proceedings of the 2002 American Control Conference (IEEE Cat. No.CH37301)*, volume 5, pages 3644–3649, 2002.

[13] Roberto Bunge and Ilan Kroo. Automatic Spin Recovery with Minimal Altitude Loss. In *2018 AIAA Guidance, Navigation, and Control Conference*, AIAA SciTech Forum. American Institute of Aeronautics and Astronautics, 1 2018.

[14] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing Function Approximation Error in Actor-Critic Methods. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR, 2018.

[15] C Liu, X Xu, and D Hu. Multiobjective Reinforcement Learning: A Comprehensive Overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398, 2015.

[16] Andrew Y Ng, Daishi Harada, and Stuart J Russell. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 278–287, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[17] Kyunghwan Cho, Jinok Shin, and Taeyong Kuc. Design of quadrotor controller for emergency situation using Xplane. In *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 311–314, 2015.

[18] G He, L Yu, S Jia, and X Wang. Simulation verification of Flight Control of a tilt tri-rotor UAV Using X-plane. In *2020 39th Chinese Control Conference (CCC)*, pages 7008–7013, 2020.

[19] A Bittar, H V Figuereido, P A Guimaraes, and A C Mendes. Guidance Software-In-the-Loop simulation using X-Plane and Simulink for UAVs. In *2014 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 993–1002, 2014.

[20] A Khalid, K Zeb, and A Haider. Conventional PID, Adaptive PID, and Sliding Mode Controllers Design for Aircraft Pitch Control. In *2019 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–6, 2019.

[21] M Rani. Modeling of an Adaptive Controller for an Aircraft Roll Control System using PID, Fuzzy-PID and Genetic Algorithm. 2016.