應用於隨機干擾飛行環境之四軸無人機強化學習控制器設計

劉俊宏[2]、葉舜斌[1]、王昱健[1]、賴威霖[1]、沈尚錡[1]、丁澤安[1]、朱力民[1]*

摘要

現今多數四軸無人機採用比例-積分-微分(Proportional-Integral-Derivative, PID)控制器，而 PID 參數常需依靠較為耗時之經驗法則進行調整，且針對特定不同飛行環境(如高度差變化量過大)，原調整後之 PID 參數亦須再次重新設計與調整，不然將導致 PID 控制器無法發揮最佳之控制性能。近來研究指出，機器學習領域中之強化學習技術可以解決高複雜度系統問題，透過接收相同飛行環境之不同誤差回饋持續進行學習，藉以獲得最佳化決策。故本研究著重於開發強化學習技術並應用至四軸無人機之姿態控制，並透過使用近端策略優化(Proximal Policy Optimization, PPO)演算法設計無人機之強化學習控制器，提高該控制器之飛行環境適應能力。本強化學習控制器與 PID 控制器於不同目標高度，將無具有及具有外部隨機干擾之情況進行控制性能分析比較，結果發現強化學習控制器之控制性能(含暫態及穩態響應)皆優於 PID 控制器，故本研究初步結論為無論在面對有無外部隨機干擾之飛行環境中，訓練完成之強化學習控制器具備更佳之環境適應性與控制能力。

關鍵字：四軸無人機、比例-積分-微分控制器、強化學習控制器、近端策略優化演算法、隨機干擾環境

[2] 劉俊宏，國立臺北大學電機工程系 助理教授。Email: chliuzzh@mail.ntpu.edu.tw
[1] 葉舜斌，國立臺東大學綠色與資訊科技學士學位學程 學生。Email: 10922127@gm.nttu.edu.tw
[1] 王昱健，國立臺東大學綠色與資訊科技學士學位學程 學生。Email: 10922123@gm.nttu.edu.tw
[1] 賴威霖，國立臺東大學綠色與資訊科技學士學位學程 學生。Email: 10822112@gm.nttu.edu.tw
[1] 沈尚錡，國立臺東大學綠色與資訊科技學士學位學程 學生。Email: 10822119@gm.nttu.edu.tw
[1] 丁澤安，國立臺東大學綠色與資訊科技學士學位學程 學生。Email: andy856996@gmail.com
[1] 朱力民(通訊作者)，國立臺東大學綠色與資訊科技學士學位學程 教授。Email: lmchu@nttu.edu.tw

Design of Reinforcement Learning Controller for Quadcopter in Flight Environment with Random Disturbance

Chun-Hung Liu[2], Shun-Pin Yeh[1], Yu-Chien Wang[1], Wei-Lin Lai[1], Shang-Chi Shen[1], Ze-An Ding[1], Li-Ming Chu[1]*

Abstract

Most quadrotors adopt Proportional-Integral-Derivative (PID) controllers, that requires time-consuming empirical tuning of PID parameters. Moreover, the parameters need to be redesigned and readjusted for various flight environments to achieve optimal control performance, especially when desired altitude is changed. Recently, some studies have demonstrated that reinforcement learning (RL) in the field of machine learning can address highly complex problems for the system. RL adopts continuous learning through feedback of different errors under the same flight environment to obtain the optimal decision for the system. Therefore, this study focuses on developing RL technology and applying it to an attitude control of quadcopters. In addition, a Proximal Policy Optimization (PPO) algorithm is used to design a RL controller for the quadcopter to enhance control performance for various flight environments. The performance of the RL controller is compared to that of the PID controller in various target altitudes without and with adding external random disturbances. The simulation results indicate that the RL controller performs better than the PID controller in terms of control performance, including transient and steady-state responses. This study preliminarily concluded that as compared with the well-tuned PID controller, this well-trained RL controller is able to have better environmental adaptability and control capability in various flight environments.

Keywords: Quadcopter, PID controller, Reinforcement Learning Controller, Proximal Policy Optimization Algorithm, Flight Environment with Random Disturbance

[2]Chun-Hung Liu, Assistant Professor, Department of Electrical Engineering, National Taipei University E-mail: chliuzzh@mail.ntpu.edu.tw

[1]Shun-Pin Yeh, Student, Interdisciplinary Program of Green and Information Technology, National Taitung University. E-mail: 10922127@gm.nttu.edu.tw

[1]Yu-Chien Wang, Student, Interdisciplinary Program of Green and Information Technology, National Taitung University. E-mail: 10922123@gm.nttu.edu.tw

[1]Wei-Lin Lai, Student, Interdisciplinary Program of Green and Information Technology, National Taitung University. E-mail: 10822112@gm.nttu.edu.tw

[1]Shang-Chi Shen, Student, Interdisciplinary Program of Green and Information Technology, National Taitung University. E-mail: 10822119@gm.nttu.edu.tw

[1]Ze-An Ding, Student, Interdisciplinary Program of Green and Information Technology, National Taitung University. E-mail: andy856996@gmail.com

[1]Li-Ming Chu (Corresponding Author), Professor, Interdisciplinary Program of Green and Information Technology, National Taitung University. E-mail: lmchu@nttu.edu.tw

壹、前言

由於四軸無人機技術的不斷發展，現今四軸無人機已廣泛應用於各種領域，例如攝影、測量、農業、運輸等。高度控制系統為四軸無人機控制器核心之一，對於升降時的穩定性和精準度影響非常大，且四軸無人機在飛行中，常會面臨著來自陣風等外部環境因素的干擾，對於飛行控制系統是一項挑戰。隨著強化學習的崛起及技術的快速發展，越來越多的領域開始將其應用到不同課題中，其中也為四軸無人機的高度控制問題提供了新的解決方法。相較於傳統的 PID 控制方法(Salih et al., 2010)，需要事先設定控制器參數，且對於複雜的環境往往難以獲得最優的控制效果；強化學習則通過在環境中與四軸無人機進行交互學習獲取最優的控制策略，進而優化高度控制的效果，對於複雜且有外部干擾之環境會有更好的適應能力。本研究著重探討使用強化學習中近端策略優化算法(Proximal Policy Optimization, PPO)(Schulman et al., 2017)做為四軸無人機高度控制器，在不同目標高度且加入隨機陣風干擾，比較並評估使用傳統 PID 控制器與本研究 PPO 控制器於高度控制之性能表現。

貳、文獻探討

一、四軸無人機之模擬環境設置

本研究使用 MATLAB/Simulink(Kaplan et al., 2019)建置模擬環境及設計控制器，並使用鸚鵡無人機(Parrot Mambo mini drone)作為模擬之四軸無人機數學模型。此款四軸無人機與 MATLAB/Simulink 具高度相關與相容性，模組化的設計方式可以提高修改和建立控制系統的效率，有利於本研究數學建模及模擬。該系統的採樣時間(sampling time, T)為 0.005 秒 (T＝0.005 (s))。本研究模擬環境大致上分成四個區塊，如 Figure 1 所示。

(一) 感測器 (Sensor)：負責整理機體傳出後的數據，並回傳給控制系統。

(二) 飛行控制系統 (Fight Control System, FCS)：主要負責整理感測器回傳的訊號資料，並與目標進行比對換算成指令對機體進行控制。

(三) 環境 (Environment)：提供模擬環境的相關環境數據，如大氣壓力、密度等。

(四) 機體 (Airframe)：接收來自飛行控制系統傳出的指令訊號，計算出機體受到物理量及環境因素影響後的位置與角速度。

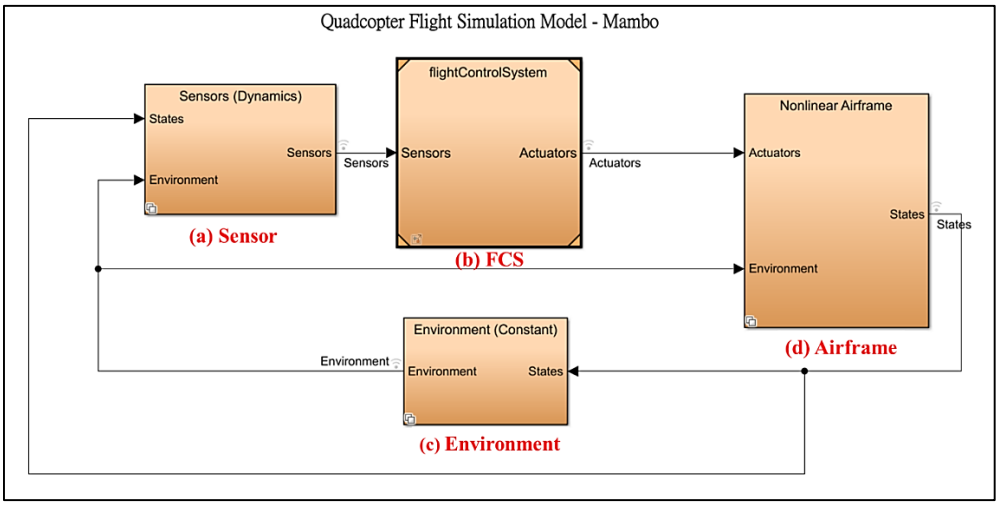

Figure 1. *模擬環境建構圖，含(a)感測器、(b)飛行控制系統、(c)飛行環境、(d)機體*

二、比例-積分-微分控制(PID)

PID 控制器(Salih et al., 2010)是一種閉迴路控制系統，因其簡單易實現且成本較低，因此本研究選擇使用 PID 控制器來進行比較。PID 四軸無人機高度控制器，透過調整三項增益值 K_p、K_i 和 K_d 控制四軸無人機的總輸出力道，以控制各馬達的轉速。PID 控制器中之比例控制項(P)用於控制當前誤差，雖可以有效控制誤差，但不利於限制震盪；積分控制項(I)用於參考過去累積的誤差與消除穩態誤差；微分控制項(D)則用於預測未來的誤差並減緩震盪。

(一) 比例控制項(P)：比例控制項的作用是使系統趨近目標值，將當前誤差與比例增益常數K_p相乘，產生一個成比例的控制量，使系統能夠快速地接近期望目標。但如果比例增益過大，會使輸出過大造成系統不穩定；反之，若比例增益過小，則控制器較不敏感。

(二) 積分控制項(I)：積分控制項負責消除穩態誤差，透過累積的誤差乘以積分增益常數K_i產生一控制量，加速系統趨近設定值的過程，並且消除比例控制器無法解決的穩態誤差。但若積分增益設計過大，無人機趨近設定值的速度過快，可能會導致過衝與震盪的情形。

(三) 微分控制項(D)：可以控制系統避免過衝或震盪，根據預測的誤差變化率和一個微分增益常數K_d相乘產生微分的控制量，減少安定時間及提高系統穩定性，微分控制項通常會加上一個低通濾波器來限制高頻增益及雜訊但需要避免放大雜訊與干擾。

PID 的方程式如下式(1)，通過計算輸入數據和理想值之間的誤差 $e(t)$，並根據誤差與累加誤差計算出更好的控制輸出，以改進系統的回饋，使系統維持在理想值上，控制的更加精確穩定。Ziegler–Nichols (ZN)方法(Ziegler & Nichols, 1942)是常用的 PID 控制器參數設定方法，特別適用於初始設定難以找到適合 PID 值之情況，以系統化有效率的方式獲得一組參考用的 PID 初始值。ZN 方法中獲得的參考值，仍然需要依據無人機的飛行環境需求進行調整以獲得最佳值。本研究使用 ZN 方法找出一組 PID 參數，並依照飛行狀況進行人工調整以獲得最佳 PID 值。

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)d\tau + K_d \frac{d}{dt}e(t) \tag{1}$$

三、強化學習(RL)

強化學習(Reinforcement learning, RL)(Wiering & Van Otterlo, 2012)為機器學習算法之一。強化學習中我們不需要標註標籤或提供準確的輸入輸出，主要藉由環境、代理(Agent)和回饋這三個元素，讓代理在環境中不斷地進行嘗試與學習，透過在線規劃(Online Planning)與環境交互訓練回饋找到最優的行為策略，以獲取最大化的期望總獎勵。對於非線性系統具有很好的適應性，可以解決傳統線性控制器無法處理的高維度複雜環境。常見的強化學習算法包括 Q-Learning (Jin et al., 2018)、Policy Gradient (Grondman et al., 2012)、PPO 等。本研究選擇 PPO 算法進行四軸無人機高度控制器設計，PPO 通過近端方法對策略進行優化，保證每次策略更新的幅度不會太大以克服過度探索策略和收斂速度緩慢等問題，具有高效且穩定等優點。

四、近端策略最佳化(PPO)

PPO (Schulman et al., 2017)主要針對策略梯度（Policy Gradient）的改進 (Engstrom et al., 2020)以確保每次策略更新的幅度有所限制，從而解決了在傳統策略梯度算法中過度探索策略和收斂速度緩慢等問題，以提高效能及增加穩定。PPO 的最終目標是通過最大化期望總獎勵來學習最佳策略，使代理能夠在給定的環境中控制系統達到預定目標。為了找到最佳策略，PPO 透過與環境的交互迭代，利用過去的數據做為經驗，集合成訓練集，並使用 CLIP（Conservative Loss Improvement for Policy）方法來限制策略更新的幅度。CLIP 方法通過對更新前後的策略分布進行限制，使用重要性採樣的技術來估計策略更新後的行動機率分佈與之前的行動機率分佈之間的差異，通過將這種差異與優勢函數（Advantage Function）相乘，可以計算出策略更新帶來的長期回報，避免大幅度更新策略所帶來的不穩定性及不良影響。PPO 採用的馬可夫決策過程(Markov Decision Process, MDP)，MDP 是強化學習算法中常用的數學框架，可用於建模具有部分隨機性和部分可控制性的序列決策問題。MDP 描述代理與環境之間的交互過程，代理根據策略在不同的狀態下選擇動作，環境根據動作給出新的狀態和即時獎勵，並根據狀態和獎勵更新策略以實現策略學習。MDP 主要由狀態集合（S, State set）、動作集合（A, Action set）、狀態轉移函數（P, State transition function）、獎勵函數（R, Reward function）和折扣因子（γ, Discount factor）組成。狀態集合代表代理在任務中所有可能出現的狀態；動作集合是指代理可以執行的所有動作；狀態轉移函數代表代理執行一個動作後，它的狀態轉移到下一個狀態的機率分布，當代理在狀態 s 執行動作 a 後，下一個狀態為 s'的機率，表示為 $\mathrm{P}(s'|s,a)$；獎勵函數則是指代理在任務中獲得的即時獎勵，表示為 $\mathrm{R}(s,a,s')$，當代理在狀態 s 執行動作 a 後，進入下一個狀態 s'時獲得的即時獎勵;折扣因子是指未來的獎勵價值應該比即時的獎勵價值低一些，γ 的值介於 0 和 1 之間，用於對長期與即時回報進行折扣。MDP 的數學框架可以應用於機器人控制、人工智慧遊戲控制等不同序列決策問題。

$$\mathrm{J}_{PPO2}^{\theta^k}(\theta) \approx \sum_{st,at} \min\left(\frac{p_\theta(a_t \mid s_t)}{p_{\theta^k}(a_t \mid s_t)} A^{\theta^k}(s_t, a_t), clip\left(\frac{p_\theta(a_t \mid s_t)}{p_{\theta^k}(a_t \mid s_t)}, 1-\varepsilon, 1+\varepsilon\right) A^{\theta^k}(s_t, a_t) \right) \quad (2)$$

(一) 網路架構：PPO的網路架構是融合策略梯度和價值函數 (cost function) 所建構的，其主要由Actor和Critic兩個部分組成，Actor-Critic網路解決策略梯度算法中的梯度變化問題。Actor學習一個策略，根據當前狀態的觀察值，通過學習權重參數來生成對應的動作機率分佈，以最大化期望獎勵；Critic則學習一個價值函數，根據當前狀態的觀察值，通過學習權重參數來估計每個動作的期望回報值，以最大化該狀態下行動的期望獎勵。而在四軸無人機PPO控制器中，Actor負責學習如何在當前環境中輸出四軸無人機最佳的控制力道，以最大化獎勵達到穩定飛行；而Critic則負責估計Actor所選擇策略的優劣程度，通過對四軸無人機飛行狀態的觀察，將其轉換為動作的機率分佈。

(二) 獎勵函數：用來評估代理在環境中採取行動是否好壞的函數。它可以定義每步所產生的獎勵或懲罰，並根據這些獎勵與懲罰來引導代理學習到正確的策略。透過與環境互動，代理會獲得狀態和獎勵，並依據策略網路產生相對應的決策動作。在本研究中以鼓勵控制器達到目標高度，並對於過衝或墜落的行為採取懲罰，根據特定條件設計出獎勵函數，並證實此獎勵函數能夠幫助代理收斂找到最佳策略。

五、性能指標

性能指標(Marzaki et al., 2015)是用來評判控制系統性能優劣的重要工具。其中，誤差積分準則是以系統輸出值與期望值誤差的函式積分作為性能指標，常見的包括平方誤差積分 (Integral Squared Error, ISE) (如式 3 所示)、絕對誤差積分 (Integral of Absolute Error, IAE) (如式 4 所示) 以及時間絕對誤差積分 (Integral Time Absolute Error, ITAE) (如式 5 所示)。不同的性能指標針對不同的情況提供控制系統解決方案。ISE 著重於系統之快速收斂並抑制最大誤差，因此使用該指標設計的控制系統通常具有較快的反應速度，但穩定度相對較差。IAE 強調系統的穩定性，負責抑制小誤差的狀況，並著重於判斷系統在穩態時的誤差，因此可以有效評估控制系統在穩態時的穩定性。ITAE 則注重於系統的調整時間，通常被用來評估系統的動態響應性能。

$$\mathrm{ISE} = \int_0^\infty \left[e(t)\right]^2 dt \tag{3}$$

$$\mathrm{IAE} = \int_0^\infty \left|e(t)\right| dt \tag{4}$$

$$\mathrm{ITAE} = \int_0^\infty t\left[e(t)\right] dt \tag{5}$$

參、研究方法

一、模擬環境及四軸無人機模型建置

本研究於 MATLAB 中匯入 Quadcopter Project 模型，並使用 Simulink 建立 Parrot Mambo mini drone 之飛行模型及模擬環境，如 Figure 1。

二、PID高度控制器

(一) 使用ZN法則調整合適之PID：本研究使用ZN調整方法中的閉環調諧方法。根據Table 1中規則，先固定其積分增益與微分增益值為0，調整比例增益值使得控制器輸出值為穩定震盪，此時的P值為K_u，根據震盪週期T_u計算出臨界增益值K_u，即可得到一組參考用的PID值。

(二) 優化ZN-PID之參數：由於ZN法則計算結果會導致系統的響應過於激烈及不穩定，因此本研究依照Table **2**進行人工調整使ZN-PID控制器效能提升，調整過後，值分別為K_p=0.57、K_i=0.01及K_d=0.27。

Table 1. *ZN方法調整表(Ziegler & Nichols, 1942)*

Control Type	K_p	K_i	K_d
P	$0.5K_u$		
PI	$0.45K_u$	$0.54K_u/T_u$	
PID	$0.6K_u$	$1.2K_u/T_u$	$0.075K_uT_u$

Table 2. *PID與暫態穩態之響應關係表(Li et al., 2006)*

	上升時間 (Rise Time)	過衝 (Overshoot)	穩定時間 (Settling time)	穩態誤差 (Steady-State Error)	穩定性 (Stability)
P 增加	減少	增加	小幅增加	減少	降低
I 增加	小幅減少	增加	增加	大幅減少	降低
D 增加	小幅減少	減少	減少	小幅變化	提升

三、PPO控制器

使用 Simulink 結合 RL agent 模塊建構 PPO 控制器，如 Figure 2。由模擬感測器中所測得機體當前高度及加速度作為 Actor-Critic 網路層輸入訊號，網路層使用完全連階層搭配 Rectified Linear Unit (ReLU) 激勵函數組合而成。本研究訓練以 1 公尺為目標高度，飛行 10 秒，超參數 (Hyperparameter)設定，Actor 的學習率 (Learning Rate)為 4×10^{-6}、Critic 的學習率 (Learning Rate)為 2×10^{-5}、熵損失 (Entropy Loss Weight)為 0.1 及批量大小 (Mini-batch Size)為 20，以上超參數決定整體訓練時速率和效果。

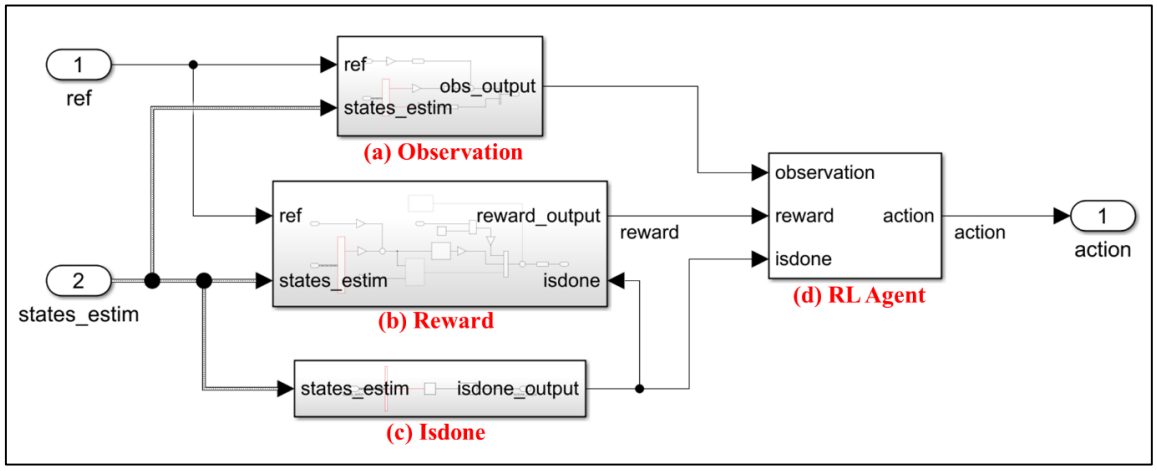

Figure 2. *PPO控制器建構圖[(a)觀察、(b)獎勵函數、(c)重製條件、(d)RL agent模塊]*

四、比較及分析數據

在模擬環境中加入外部隨機干擾(Wang et al., 2019)，以目標高度分別為 1.0、2.5 及 2.0 公尺連續飛行 30 秒，比較 ZN-PID 控制器及 PPO 控制器在具有外部隨機干擾及無干擾下飛行狀態。由 Z 軸方向產生持續 1 秒隨機大小之力，並每秒變換力的大小與方向，其用意為模擬戶外環境中不固定之自然風。本研究設計將干擾之力轉換成對四軸無人機造成之位移，將此模塊加入 Figure 1 (c)中，每秒位移大小如 Figure 3。最後藉由暫態響應各項數據搭配性能指標，評估兩種控制器在不同高度及干擾下之性能。

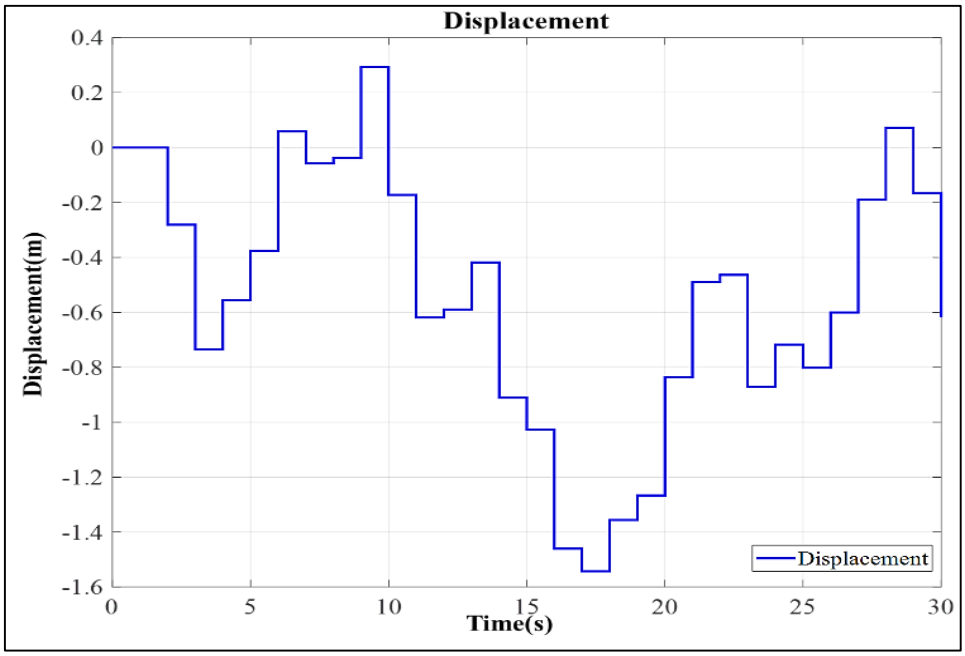

Figure 3. *加入干擾力轉換成對四軸無人機造成之高度位移影響*

肆、結果與討論

本研究中 ZN-PID 及 PPO 控制器均以 1.0 公尺為目標高度進行訓練及設計，測試高度分別為 1.0 公尺、2.5 公尺及 2.0 公尺，以此測試面對不同目標高度時，控制器表現。Figure 4 為模擬無干擾的環境可觀察到不論是上升還是下降 PPO 控制器都沒有過衝的現象，飛行也都相當平穩；ZN-PID 有明顯的過衝，且在上升高度 1.5 公尺時無法在 10 秒內收斂，下降時也有過衝的現象。Table **3** 數據呈現上升 1.0 公尺時 PPO 比 ZN-PID 控制器節省了 3.28%的時間，在過衝方面 PPO 只有過衝 0.01 公尺，比 ZN-PID 的 0.22 公尺抑制多達 95.45%。Figure 5 在飛行時加入外部隨機干擾模擬自然風，使模擬環境最符合實際飛行狀況。由 Figure 3 可知測試干擾的位移為每秒變換之 0.3 公尺至負 1.5 公尺。Figure 5 中可以看出 PPO 振幅明顯較小，Table **3** 中各項性能指標 PPO 均優於 ZN-PID 控制器，以 ITAE 作為判對動態響應的性能指標，PPO 較 ZN-PID 改善了 23.62%。

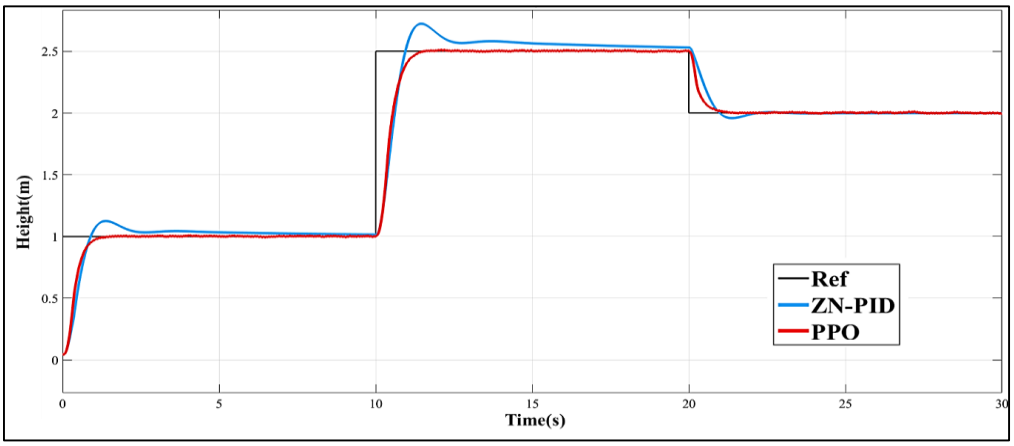

Figure 4. *PPO及ZN-PID在無干擾時飛行圖(高度分別為1.0m、2.5m及2.0m)*

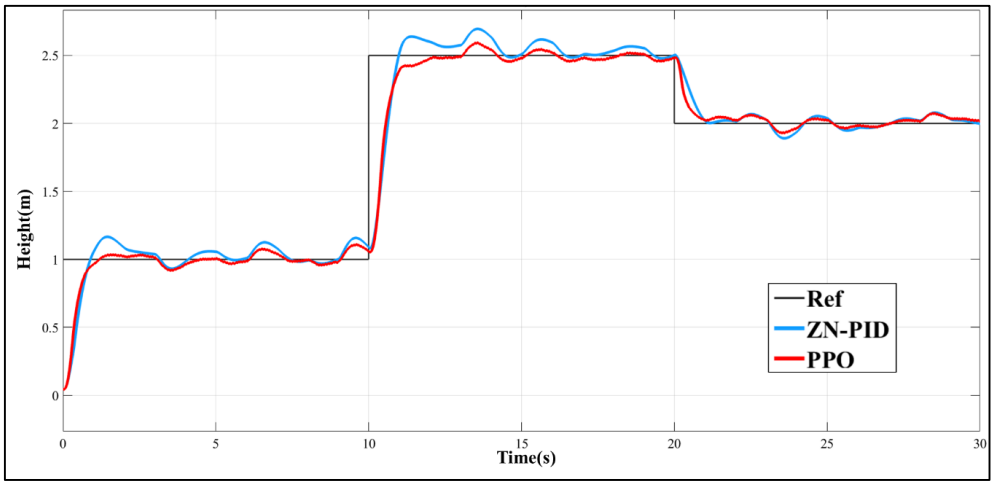

Figure 5. *PPO及ZN-PID在干擾時飛行圖(高度分別為1.0m、2.5m及2.0m)*

Table 3. *ZN-PID與PPO控制器於有無環境干擾之性能指標及改善程度比較*

	Benchmark ZN-PID	PPO	Improvement (%)
Without Disturbance			
ITAE (m・s)	4817.47	2426.32	49.63 %
IAE (m)	2.45	1.37	44.08 %
ISE (m^2)	1.29	1.07	17.05 %
1.0 m Rise Time (s)	0.61	0.59	3.28 %
2.5 m Rise Time (s)	10.47	10.5	-0.29 %
Peak (m)	0.22	0.01	95.45 %
With Disturbance			
ITAE (m・s)	6501.74	4965.88	23.62 %
IAE (m)	2.89	2.19	24.22 %
ISE (m^2)	1.33	1.14	14.29 %
1.0 m Rise Time (s)	0.61	0.59	3.28 %
2.5 m Rise Time (s)	10.49	10.53	-0.38 %
Peak (m)	0.19	0.09	52.63 %

伍、結論

本研究藉由三種不同目標高度測試 PPO 高度控制器之適應能力，並使用傳統 PID 控制器搭配 ZN 法則作為對照組，觀察 PPO 控制器的優劣勢。兩種控制器皆以 1 公尺為基準設計及訓練，PID 控制器在上升距離較大時，會有嚴重過衝及無法收斂等問題；PPO 控制器在不同高度不論是過衝或響應速度皆維持一定水準的表現，其反映了 PPO 面對高度變化具備強大之適應能力。加入隨機干擾的模擬環境中，觀察 ZN-PID 與 PPO 控制器可以看出 PPO 控制器的性能指標表現更好，在動態響應與過衝方面，PPO 控制器的表現皆優於 ZN-PID 控制器，因此在面對有外在干擾的飛行環境中，PPO 控制器具有更佳的適應性與控制能力。結果顯示 PPO 控制器在飛行控制系統上相較於傳統 PID 控制器有更好的性能表現，未來將繼續研究四軸無人機 PPO 六軸控制器並應用於實際飛行中。

陸、致謝

本研究感謝國科會計畫(NSTC 109-2222-E-305-002-MY3)高速運算電腦運算之支持及教育部高等教育深耕計畫之支持。

柒、參考文獻

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., & Madry, A. (2020). Implementation matters in deep policy gradients: A case study on ppo and trpo. *arXiv preprint arXiv:2005.12729*.

Grondman, I., Busoniu, L., Lopes, G. A., & Babuska, R. (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 1291-1307.

Jin, C., Allen-Zhu, Z., Bubeck, S., & Jordan, M. I. (2018). Is Q-learning provably efficient? *Advances in neural information processing systems*, *31*.

Kaplan, M. R., Eraslan, A., Beke, A., & Kumbasar, T. (2019). Altitude and position control of parrot mambo minidrone with PID and fuzzy PID controllers. 2019 11th International Conference on Electrical and Electronics Engineering (ELECO),

Li, Y., Ang, K. H., & Chong, G. C. (2006). PID control system analysis and design. *IEEE Control Systems Magazine*, *26*(1), 32-41.

Marzaki, M. H., Tajjudin, M., Rahiman, M. H. F., & Adnan, R. (2015). Performance of FOPI with error filter based on controllers performance criterion (ISE, IAE and ITAE). 2015 10th Asian control conference (ASCC),

Salih, A. L., Moghavvemi, M., Mohamed, H. A., & Gaeid, K. S. (2010). Flight PID controller design for a UAV quadrotor. *Scientific research and essays*, *5*(23), 3660-3667.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Wang, B. H., Wang, D. B., Ali, Z. A., Ting Ting, B., & Wang, H. (2019). An overview of various kinds of wind effects on unmanned aerial vehicle. *Measurement and Control*, *52*(7-8), 731-739.

Wiering, M. A., & Van Otterlo, M. (2012). Reinforcement learning. *Adaptation, learning, and optimization*, *12*(3), 729.

Ziegler, J. G., & Nichols, N. B. (1942). Optimum settings for automatic controllers. *Transactions of the American society of mechanical engineers*, *64*(8), 759-765.