Journal of Aeronautics and Space Technologies

15(2) (2022) 94-107

Journal Homepage: jast.hho.msu.edu.tr

RESEARCH ARTICLE

Deep Reinforcement Learning-based Cooperative Survivability Maximization for a UAV Fleet on an Air-to-Ground Mission

Havadan Karaya Görev Yapan İHA Filosu için Derin Pekiştirmeli Öğrenme Tabanlı İşbirlikçi Beka Optimizasyonu

Barış Başpınar[1*]

[1] *Istanbul Technical University, Department of Aeronautical Engineering, 34467 Istanbul, Turkey,*

Received: *February 22, 2022* ***Revised:*** *June 29, 2022* ***Accepted:*** *July 04, 2022*

Abstract

This study focuses on the cooperative strategy development of a UAV team that operates in a hostile environment in which the radar and weapon systems try to track and eliminate them. To simulate the hostile defense system, we present Markov models that generate the detecting and tracking probabilities of a radar system, and calculate the multiple-shot survivability of air vehicles that fly within the hostile environment. A cooperative strategy development procedure is presented based on proximal policy optimization algorithm, which is a deep reinforcement learning method. It is shown that the UAV team can develop cooperative strategies by exploiting enemy's weakness to maximize team survivability in an air-to-ground mission after training with the proposed reinforcement learning scheme.

Keywords: Aerospace, Deep Reinforcement Learning, Air-to-Ground Survivability, Proximal Policy Optimization, Cooperative Strategy Development, UAV Team.

Özet

Çalışma kapsamında hasıma ait radar ve silah sistemlerinin bulunduğu bir ortamda operasyon yapan İHA takımının işbirlikçi strateji geliştirmesine odaklanılmıştır. Hasım savunma sisteminin benzetimini yapmak için Markov modelleri sunulmuştur. İlgili modeller radar sisteminin tespit ve takip olasılıklarının üretebilmekte ve hasım ortamında uçuş yapan hava araçlarının çoklu atış bekalarını hesaplayabilmektedir. Bir derin pekiştirmeli öğrenme metodu olan Proksimal Politika Algoritması vasıtasıyla bir işbirlikçi strateji geliştirme yöntemi sunulmuştur. Önerilen pekiştirmeli öğrenme yapısıyla eğitimin gerçekleştirilmesi ardılında İHA takımının rakibin zayıflığını kullanarak takım bekasını en iyilerken işbirlikçi strateji geliştirebildiği gösterilmiştir.

Anahtar Kelimeler: Havacılık, Derin Pekiştirmeli Öğrenme, Havadan Karaya Beka, Proksimal Politika Optimizasyonu, İşbirlikçi Strateji Geliştirme, İHA Takımı.

1. INTRODUCTION

Air-to-ground combat is a complex task due to the interaction of multiple decision-makers. It contains numerous elements that have an impact on the decision process of the commander. In this process, the primary goal is to achieve the defined mission while withstanding the hostile environment. The ability of an aircraft or a UAV (unmanned aerial vehicle) team to withstand and avoid the hostile environment is defined as combat survivability, which is presented as a probability [1].

*Corresponding Author
E-mail: baspinarb@itu.edu.tr

The survivability assessment begins with a mission-threat analysis as a direct consequence of the correlation between the threat capability and combat survivability. This assessment should contain the definition of the mission, the air vehicles' abilities, and the threats' capabilities. Based on the developed models for the natural components of the air-to-ground combat, an optimized strategy can also be developed to accomplish the mission.

There are several studies in the literature associated with survivability assessment and mission planning procedures in a hostile environment. Using the theory of the stochastic duel, an analytical model that calculates the survival probability for the one-on-one engagement between an aircraft and a weapon system is developed in the study [2]. The aircraft's survivability is also analyzed by focusing on the vulnerability of the aircraft against a fragmenting warhead threat [3]. The study [4] assesses the survivability using a Markov model with five-state variables. These studies focus only on aircraft survivability modeling without considering mission planning. There are also studies focused on path planning in a hostile environment. The majority of these studies use threat cost functions related to the distance between the air vehicle and threat to ensure that the aircraft avoids threats as much as it can when the aircraft operates inside a danger zone [5-8]. The study [5] focuses on route planning via evolutionary algorithms by defining a lumped objective function that aims to minimize the violation of danger zones. In addition to the evolutionary algorithms, the predictive model control is also used to perform the path planning task in a threat region [6]. Instead of minimizing the violation of the danger zones, this study proposes a threat cost that is inversely proportional with r^4 by inspiring the radar equation, where r is the range to target. The study [7] utilizes the Voronoi diagram to generate the route with a similar objective function. The study [8] also uses a similar approach to present the radar and weapon area costs, and solves the path planning problem via A* algorithm. These kinds of studies tend to use simplified threat models. Their threat models have no dynamics and are incapable of assessing the multi-shot survivability. Furthermore, it can be sometimes hard to use this threat modeling approach for multi-UAV mission planning. Assume that the radar can only track an air vehicle at a time. It will be not possible to use the aforementioned approach in this scenario. In our previous study [9], we proposed a more complex threat modeling approach in which radar and weapon systems have dynamics and are capable of calculating the multi-shot survivability. In this study, we have extended our research by focusing on cooperative strategy development for a UAV team in a hostile environment via proximal policy optimization in which the problem is evaluated in continuous spaces. Instead of focusing on a single aircraft, the proposed approach evaluates the team members together and generates a collaborative strategy to maximize the team survivability on an air-to-ground mission by assessing the multi-shot survivability.

Firstly, the overall method is presented. Then, the models that simulate the air defense system are explained in detail. These models correspond to the radar and weapon systems inside the environment of the enemy and are used to calculate the survivability of an air vehicle by evaluating its present state and the elements in the environment. Next, the equations of motion that are used to present the UAV dynamics are described. Then,

the cooperative strategy development approach based on deep reinforcement learning is presented.

The rest of this paper is organized as follows. Section 2 explains the methodology that contains the modeling of the air defense system, UAV dynamics, and the details of the deep reinforcement learning approach. The concept and validity are demonstrated through simulation results in Section 3. Then, concluding remarks are provided in Section 4.

2. METHODOLOGY

In this study, a cooperative strategy generation procedure based on deep reinforcement learning is proposed. Reinforcement learning (RL) is mainly about learning through experience [10]. During the training process, decision-making agents interact with their environments and receive positive or negative feedback based on their actions. In this way, they try to generate a strategy that maximizes the cumulative reward. The state of the environment changes according to the agents' actions. This process is a Markov Decision Process when the agents can fully observe the state. The MDPs can be solved using RL methods. RL algorithms define how an agent updates its policy according to its experience. The training process is based on fast-time simulations. During the training, the agent tries to complete the defined task by maximizing its total reward and updates its policy according to the defined objective. The policy is updated for each episode and converges to an optimized solution at the end of the process. In deep reinforcement learning, the policy is presented with a neural network, and the weights of the neural network are adjusted during the training to optimize the policy.

The interaction between the agents and environment can be assessed via simulations. A virtual air defense system is presented and the UAV dynamics are presented via differential equations in this study. Then, using the developed models, a simulation environment is created. The proposed RL algorithm based on proximal policy optimization is utilized in the simulation environment to train the policy network to generate an optimized strategy for cooperative survivability maximization in an air-to-ground mission.

In the rest of this section, the models for air defense system, UAV dynamics, and learning-based cooperative strategy generation process are described sequentially.

2.1. Air Defense System

An air defense system contains several threat elements with different capabilities and characteristics. This system aims to outperform the enemy air forces to prevent them from reaching their goals. There are many elements such as missiles, radars, communication links, launch platforms, fire units in this system [1]. These components can be considered as the elements of either radar system or weapon system.

In an air defense system, the radar system searches the environment to find the targets. After detecting them, the system tracks the targets to provide the required information for the weapon system. The weapon system contains a fire control computer that has prior information about the weapon's abilities. It generates a fire solution using the provided target information and the weapon's parameters. When surface-to-air missiles

(SAM) are used as the main weapon, a missile is fired according to the calculated fire solution. The initial boost, mid-course, and terminal phase are the three main phases in a missile intercept. The initial boost begins without guidance, then the missile is directed to the opponent in the consecutive phase. It tries to reach the target in the terminal phase to detonate the fuse. The ground system can assist the missile during the last two phases to track the opponent and update the fire solution depending on the guidance technique. For example, the in-flight corrections are provided by the radar system to the missile in the command guidance.

In this study, we have benefited from the threat models developed in the study [9]. The details of the air defense system are presented in the rest of this subsection.

2.1.1. Mathematical Model of Radar System

The radar system attempts to detect and locate the targets. The area should be actively scanned to detect the opponents. Then, the revealed opponents should be tracked to obtain the required target information to locate them.

In this study, the radar system is presented as a Markov process as proposed in the study [9]. The states of the radar are defined as the probabilities of scanning, detecting, and tracking. The transitions between the states are modeled using exponential decays.

The following dynamical system is used to describe the transitions between the states to model the radar system:

$$\begin{aligned} \dot{p}_s &= -\lambda_{sd} p_s + \lambda_{ds} p_d \\ \dot{p}_d &= \lambda_{sd} p_s - (\lambda_{ds} + \lambda_{dt}) p_d + \lambda_{td} p_t \\ \dot{p}_t &= \lambda_{dt} p_d - \lambda_{td} p_t \end{aligned} \tag{1}$$

where λ's are the coefficients depending on the radar and target properties. p_s, p_d and p_t refer to the probability of scanning, detection probability, and probability of tracking, respectively. The initial condition of the state vector $[p_s, p_d, p_t]^T$ is $[1, 0, 0]^T$ that corresponds to the scanning of the environment when there is no target in radar coverage.

The major parameters that have a direct impact on detecting and locating a target via a radar system are the signal-to-noise ratio (SNR) and jamming-to-signal ratio (J/S). The SNR corresponds to the ratio of the received signal strength to the noise strength, and the target can be revealed when it exceeds a specific threshold. The signal power must be higher than the noise power to detect the target. The ratio is expressed as follows [11]:

$$SNR = \frac{P_T G \sigma A_e}{KTBF(4\pi)^2 R^4} \tag{2}$$

where R is the distance between the target and radar. P_T, G, σ, and A_e refer to the transmitted power, antenna gain, radar cross-section of the target, and antenna aperture area. K, T, B, and F are the parameters used to calculate the level of noise and are defined as Boltzmann's constant, temperature, equivalent bandwidth, and noise figure, respectively. The SNR is an indicator of the target detection ability of the radar and also presents the vulnerability of the radar to noise jamming.

The J/S refers to the ratio of the signal strength of the jamming to the received signal strength. It is the key indicator of jamming effectiveness. This ratio is presented as follows [11]:

$$J/S = \frac{P_J G_J}{P_T G_T} \frac{4\pi R^2}{\sigma} \tag{3}$$

where, P_J, G_J, P_T, and G_T correspond to the transmitted jamming power, jamming antenna gain, transmitted peak power, and radar antenna gain, respectively.

By considering the information provided, λ_{sd} and λ_{dt} are directly proportionate to SNR, and λ_{td} is proportionate to J/S. And, the coefficient λ_{ds} is defined as proportional to the stealth feature of aircraft that changes linearly with $1/\sigma$.

The main parameter that directly changes both SNR and J/S is the distance R between the radar and target. The rest of the quantities in the expressions depend on the radar model and target type. Hence, the transition coefficients can be described as follows:

$$\lambda_{sd} = \lambda_{dt} = \begin{cases} K_1\left(\frac{R_{max}^4}{R^4}\right) & if\ \sqrt[4]{K_1} R_{max} < R \leq R_{max} \\ 1 & if\ R < \sqrt[4]{K_1} R_{max} \\ 0 & otherwise \end{cases} \tag{4}$$

$$\lambda_{td} = K_2\left(\frac{R^2}{R_{max}^2}\right) \tag{5}$$

$$\lambda_{ds} = K_3/\sigma \tag{6}$$

where R is the range to the target. R_{max} refers to the maximum unambiguous range. $K_1, K_2, K_3/\sigma \in [0,1]$ corresponds to the coefficients that cover the rest of the radar and UAV specific parameters.

2.1.2. Tracking Mission Progress

The progress in a mission can be described as a Markov process in which the state variables are specified as processing and completed. A subprocess in the weapon system can be defined via this model. The following equations are used to describe the model:

$$\dot{p}_p = -\lambda_m p_p \tag{7}$$

$$\dot{p}_c = \lambda_m p_p \tag{8}$$

where, λ_m defines the time-dependent behavior of the mission. The initial condition of the state vector $[p_p, p_c]^T$ is defined as $[1,0]^T$. The mission status (M_{status}) is described according to the probability of completed p_c as follows:

$$M_{status} = \begin{cases} 1 & if\ p_c \geq 0.99 \\ 0 & otherwise \end{cases} \tag{9}$$

2.1.3. Survivability Model

In this study, the weapons in the air defense system are defined as SAMs. The shooting process of a SAM is described with three consecutive sub-missions that are obtaining a fire solution, launching the missile, and intercept phase. The aforementioned mission model is used to simulate each of these submissions. They are initialized when the following conditions are satisfied:

$$M_{ini}^{fire} = \begin{cases} 1 & if\ p_t \geq 0.7\ and\ |x_a - x_0| \leq R_{SAM}^{max} \\ 0 & otherwise \end{cases} \tag{10}$$

$$M_{ini}^{launch} = \begin{cases} 1 & if\ M_{status}^{fire} = 1 \\ 0 & otherwise \end{cases} \tag{11}$$

$$M_{ini}^{intercept} = \begin{cases} 1 & if\ M_{status}^{launch} = 1 \\ 0 & otherwise \end{cases} \tag{12}$$

where, $|x_a - x_o|$ is the distance between the aircraft and weapon, and p_t refers to the probability of tracking obtained from the radar system. The firing process is triggered when an enemy is inside the weapon region and the tracking probability p_t reaches the defined threshold. After completing the firing solution phase, the launching process is initiated. Then, the intercept phase follows the launching process. The completion of the last phase corresponds to a single shot by the weapon system. Then, all sub-missions are reset. A consecutive shooting process will be initiated if the target still satisfies the required conditions of the firing process.

The target's survivability is affected by each shooting process. Each completed shot reduces the probability of survival of the target. Hence, the kill probability of a single shot is computed to update the target's survivability. The killability or kill probability (P_K) is given as follows:

$$P_K = P_H P_{K|H} \tag{13}$$

where, $P_{K|H}$ and P_H are the vulnerability and susceptibility, respectively [1]. Both vulnerability and susceptibility define the inability of the target to withstand the defense system. The vulnerability is the conditional probability that the missile destroys the target given that the target has hit. The susceptibility P_H is the single shot hit probability and can be presented as follows:

$$P_H = P_{H|I} P_{I|L} P_{L|t} p_t \tag{14}$$

where, p_t is the probability of tracking obtained from the radar system. The rest of the parameters are conditional probabilities based on the missile launching and engagement process. $P_{L|t}$ is the probability of a successful launch given that the target has been tracked. $P_{I|L}$ is the probability of a successful intercept for a launched missile. $P_{H|I}$ is the hit probability of the missile given that it has been intercepted.

The single-shot killability P_K is used to calculate the aircraft's survivability under a single shot (P_S^{ss}) as follows:

$$P_S^{ss} = 1 - P_K \tag{15}$$

When the shot outcomes are evaluated independently, the multiple-shot survivability for the aircraft can be given as follows:

$$P_S^a \leftarrow P_S^a P_S^{ss} \tag{16}$$

where, P_S^a is the survival probability of the aircraft under multiple-shot. P_S^a is taken as 1 at the beginning of the simulation, and it is updated according to the expression (16) whenever a shooting process is finalized.

2.2. Aircraft Equations of Motion

A two-dimensional continuous-time model is used to simulate the motion of a fixed-wing UAV as in the study [12]. The control inputs are defined as the velocity V and roll angle ϕ of the UAV. The position (x, y) and heading angle ψ of the vehicle are described via kinematic equations depending on the control inputs. The UAV model is presented by the following set of differential equations:

$$\begin{aligned} \dot{x} &= V\cos\psi \\ \dot{y} &= V\sin\psi \\ \dot{\psi} &= \frac{g\tan\phi}{V} \end{aligned} \tag{17}$$

2.3. Learning-based Cooperative Strategy Generation

As previously mentioned, in RL, the focused problem is described as an MDP. And, an MDP is defined by the tuple (S, A, T, R). S and A are state space and action space. T: $S\ x\ A\ x\ S \rightarrow \mathbb{R}$ is the transition probability distribution. $R: S\ x\ A \rightarrow \mathbb{R}$ is the reward function.

The transition dynamics is already presented with the previously described models. Using the information obtained from the models, the state and action spaces of the MDP need to be defined. And, the reward function and RL algorithm have to be determined.

2.3.1. State and Action Spaces

The action space has to contain the inputs of the UAVs. Because the model (17) is used to simulate the motion of the agents, each UAV has two input variables, which are the speed and bank angle. These variables are normalized between -1 and 1. Then, the action space for the first UAV is defined as follows:

$$A_1 = \left[\frac{v_{max}-v_{min}}{2}v_1 + \frac{v_{max}+v_{min}}{2}, \frac{\phi_1}{\phi_{max}}\right] \tag{18}$$

where v_{min} and v_{max} are the lower and upper speed limits. v_1 and ϕ_1 are the current speed and bank angle of the UAV_1.

For a fleet of N UAVs, the action space consists of the normalized control inputs of these UAVs and is defined as follows:

$$A = \{A_1, A_2, \dots, A_N\} \tag{19}$$

The state space must contain all information regarding the mission. The state space for the first UAV is defined as follows:

$$S_1 = \left[\frac{x_1 - x_T}{L}, \frac{y_1 - y_T}{L}, \frac{\chi_{1\backslash T}}{\pi}, \frac{\psi_1}{\pi}, \frac{d_{1\backslash T}}{L}, P_S^1, p_t^1, p_d^1, \frac{d_{1\backslash th_1}}{L}, \frac{\chi_{1\backslash th_1}}{\pi}, \dots, \frac{d_{1\backslash th_M}}{L}, \frac{\chi_{1\backslash th_M}}{\pi}\right] \quad (20)$$

where (x_1, y_1) is the current position of the UAV_1, (x_t, y_t) is the center of the target/destination area. The edge length of the predefined region of operation (L) is used to normalize the distances. ψ_1 is the heading angle of the UAV_1. $d_{1\backslash T}$ and $\chi_{1\backslash T}$ are the relative distance and relative course with respect to the destination. P_S^1, p_t^1, and p_d^1 symbolize the survival probability, tracking probability, and detecting probability for the UAV_1, respectively. $d_{1\backslash th_1}$ and $\chi_{1\backslash th_1}$ describe the relative distance and relative course with respect to the M^{th} threat .

For a UAV fleet, the state space consists of the described states for all UAVs and is presented as follows:

$$S = \{S_1, S_2, \dots, S_N\} \quad (21)$$

2.3.2. Reward Function

In the focused problem, a UAV team tries to reach a target region in a hostile environment. To speed up the convergence in the training process, a reward function is defined using the distances to the destination region as follows:

$$R_{dist} = \sum_{i=1}^{N} -1.7e^{1.5\frac{d_{i\backslash T}}{L}} \quad (22)$$

where L is the edge length of the predefined simulation region used for the normalization, and $d_{i\backslash T}$ is the Euclidean distance between the current position of the i^{th} UAV and the center of the target area. So, to reduce the high negative rewards, the UAVs approach the destination area when they are far from the destination.

It is also required to avoid being detected and tracked. The corresponding reward values are generated using the threat models. For avoiding exposure, a penalty is defined as:

$$R_{radar} = \sum_{i=1}^{N} -0.5p_d^i - 1.5p_t^i \quad (23)$$

where, p_d^i and p_t^i are probabilities of detecting and tracking of the i^{th} UAV, respectively.

Whenever the survivability of a UAV decreases because of a successful shot by the missile system, a negative reward is also given according to the decrease in the survivability of the corresponding UAV as follows:

$$R_{kill}^i = -1500\left(P_{S_{old}}^i - P_S^i\right) \quad (24)$$

where $P_{S_{old}}^i$ is the survivability of the i^{th} UAV before the shot, and P_S^i is the survivability of the i^{th} UAV after the shot.

When a UAV reaches the destination area, its simulation is terminated. And, because of the mission accomplishment, a positive reward is given as follows:

$$R^i_{reach} = 1500 \tag{25}$$

Furthermore, a penalty is defined to prevent unnecessary searches in far areas. When a UAV leaves the predefined region, a negative reward is given as follows:

$$R^i_{out} = -500 \tag{26}$$

During the training, an epoch is terminated whenever an agent leaves the predefined region or is killed by a missile system. It is assumed that a UAV is killed when its survivability is smaller than 0.05. When all members of the team reach the target area, the epoch is also terminated.

2.3.3. Learning Algorithm

In this study, Proximal Policy Optimization (PPO) algorithm [13], which is a deep reinforcement learning algorithm, is used to train the UAVs for developing optimized cooperative strategies. It is an on-policy learning method in an actor-critic framework. In each iteration, the algorithm tries to maximize the following objective function:

$$L_t(\theta) = \hat{\mathbb{E}}_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)] \tag{27}$$

where θ is the vector of policy parameters, c_1, c_2 are scaling coefficients, S corresponds to the entropy bonus, and L_t^{VF} is the squared-error loss. $L_t^{CLIP}(\theta)$ is used to ensure that the policy network converges according to obtained rewards and is presented as follows:

$$L_t^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{28}$$

where epsilon is a hyperparameter, $\epsilon = 0.2$. The estimated advantage $\hat{A}_t$ corresponds to the discounted rewards minus baseline estimate. And, $r_t(\theta)$ denotes the probability ratio and is presented as follows:

$$r_t(\theta) = \frac{\pi_\theta(a_t, s_t)}{\pi_{\theta_{old}}(a_t, s_t)} \tag{29}$$

here, $\pi_{\theta_{old}}$ is the policy before the update.

The PPO algorithm is presented below. In each iteration, each of N parallel actors collect T timesteps of data, and average estimates are calculated. Then the objective L (27) is optimized using the collected information by adjusting the vector of policy parameters (θ) via stochastic gradient descent (or Adam algorithm). In this way, the policy network is trained iteratively.

Table 1. PPO algorithm.

For each episode:
For each step of episode:
Run old policy $\pi_{\theta_{old}}$ T timesteps
Compute average estimates $\hat{A}_1, \dots, \hat{A}_T$
Optimize L (27) with respect to θ, with K epochs
Update the old policy $\pi_{\theta_{old}} \leftarrow \pi_\theta$

3. IMPLEMENTATION

The capability and validity of the proposed approach are demonstrated through simulations. The threat models can generate the radar probabilities and combat survivability according to the route followed by the air vehicle, but it is necessary to define their performance parameters. In the implementation, the performance parameters of the threat models and UAVs are specified as presented in Table 2.

Table 2. Performance parameters and limits for air defense system and UAVs.

Parameter	Value	Parameter	Value
K_1	0.1	K_2	1
$P_{H\|I}$	0.7	$P_{I\|L}$	0.7
$P_{L\|t}$	1	$P_{K\|H}$	0.5
λ_m^{fire}	5	λ_m^{launch}	5
$\lambda_m^{intercept}$	2	λ_{ds}	0.2
v_{min}	$30\ m/s$	v_{max}	$40\ m/s$
ϕ_{min}	$-30°$	ϕ_{max}	$30°$

It is assumed that a radar system can only track an aircraft at a time. In the implementation, the radars will be locked to the closest enemy inside their coverage areas and track it until losing the signal. This drawback of the radar system creates an opportunity for the UAV team to explore and exploit the enemy's weakness.

Firstly, we have focused on a case study in which there are five threats in the simulation environment as shown in Figure 1. The radar areas are shown with red circles, while the weapon areas are illustrated with blue circles. And, it is assumed that there is an uncertainty associated with the center location of the threats, and this uncertainty is modelled using a Gaussian distribution. During the training, the centers of the threats are shifted randomly according to this distribution. In this way, the position uncertainties are also considered.

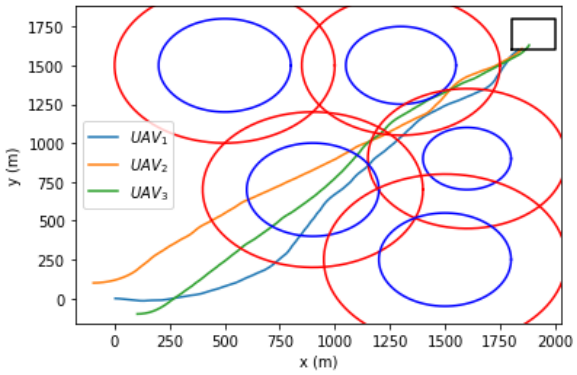


Figure 1. Developed cooperative strategy by the trained policy network for the UAV fleet in the hostile environment (Scenario 1).

A team of three UAVs is trained in the defined hostile environment to generate an optimized strategy to reach the target region behind the threat area as illustrated in Figure 1.

In Figure 1, the light blue, orange, and green paths correspond to the optimized routes for the UAV team. As shown in this figure, the optimized trajectories for the UAV team are generated successfully to reach the target area, which is illustrated as a black square. The performance of the learning algorithm during the training is presented in Figure 2. The cumulative reward of the team converges to a specific point. This figure shows that the training process is completed successfully.

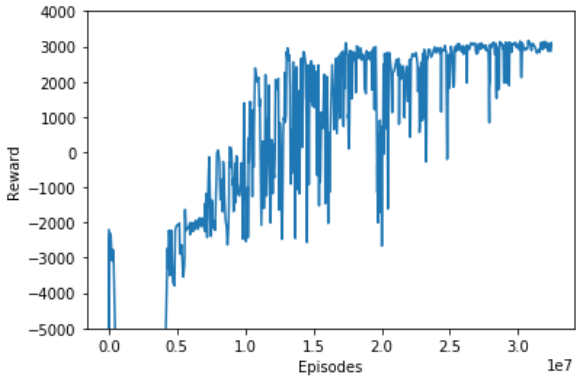


Figure 2. Episode rewards of the policy during the training process.

As presented in Figure 1, the UAVs are moving together during the operation. This strategy is developed to maximize the team survivability. The survivability of each UAV during the operation is presented in Figure 3. It is seen that only the survivability of the UAV_3 is decreasing during the operation, while the others complete the mission without a change in their survivabilities. As illustrated in Figure 4, the radars are always tracking the UAV_3 during the operation. The UAV team develops a strategy to maximize the team survivability by staying close to each other during the whole mission. When they move together, they enter the same threat areas and the radar system can only track one of them. By exploiting this strategy, the UAV team maximizes the team survivability.

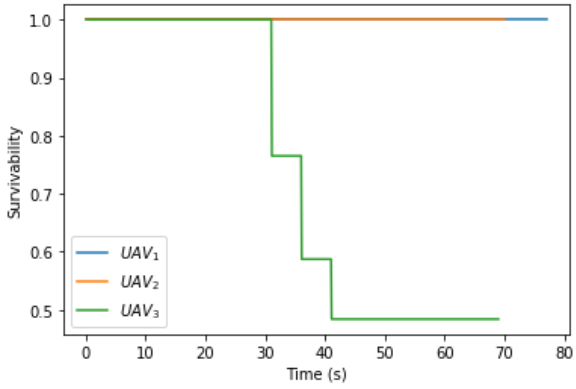


Figure 3. Survivabilities of the UAVs during the air-to-ground mission.

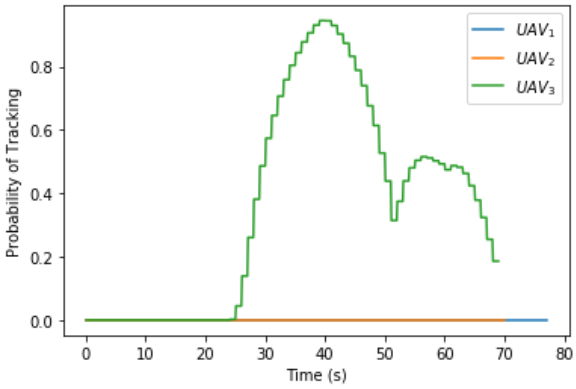


Figure 4. Tracking probabilities of the hostile radars.

In the second scenario, we have focused on a case study in which there are two UAVs in a less complicated hostile environment to illustrate that the team can be capable of completing the mission without interfering with the threats when there is a way to do that. The second scenario and the strategy of the UAV team are presented in Figure 5.

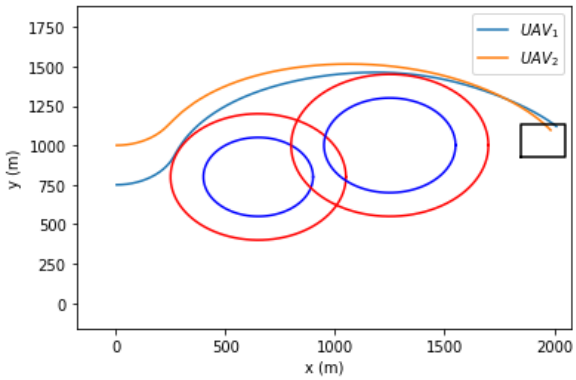


Figure 5. Developed strategy for scenario 2.

As illustrated in the simulation results in Figure 5, the team is capable of avoiding the threat regions and reaching the target area. The radars cannot detect any of the UAVs and the team completes the mission without any reduction in the radar probabilities and team

survivability. The case study shows that the UAV team can be completely away from the threat regions when this is a feasible solution.

4. CONCLUSION

In this study, an air-to-ground mission assessment platform has been developed for cooperative strategy generation for UAV teams. Threat models were presented to calculate the radar state probabilities and combat survivability in a hostile environment. Based on the proximal policy optimization, a deep RL approach has been used to generate optimized policies for UAVs. The capability and validity of the proposed approach have been demonstrated through simulations. And, it has been shown that the UAV team can explore and exploit the weakness of the opponent to achieve the defined mission and maximize the team survivability via the proposed approach.

In addition to the strategy development, the proposed platform can be used for mission and inventory assessment in multi-UAV operations. The impact of air vehicles' specific features and countermeasures on operation can also be evaluated. The performance of different air vehicles and different team configurations can be analyzed via generated measures in the proposed platform to decide on an optimized team configuration. For example, the vulnerability of an aircraft that operates inside a weapon region has an impact on the team survivability. The radar state probabilities can also be manipulated with the stealth factor because of its direct impact on λ_{sd}. To assess the survivabilities of different team configurations for a specific mission, the team survivability, total reward, and other properties obtained from the developed platform can be used. Similarly, the impact of the countermeasures such as the radar jamming that changes the radar state probabilities, chaffs and flares that affect the $P_{H|I}$ and $P_{I|L}$ on the survivability can also be assessed via the proposed platform for determining the required equipment and developing strategies. An alternative implementation can be about considering the uncertainties in intelligence during the mission planning process to generate a more robust cooperative strategy for a UAV team.

REFERENCES

[1] R. E. Ball, *The fundamentals of aircraft combat survivability analysis and design*. Reston, VA: AIAA (American Institute of Aeronautics & Astronautics), 2003.

[2] X. Wang, B.-F. Song, and Y.-F. Hu, *Analytic model for aircraft survivability assessment of a one-on-one engagement*, Journal of Aircraft **46**(1), 223–229 (2009).

[3] H. E. Konokman, A. Kayran, and M. Kaya, "Analysis of aircraft survivability against fragmenting warhead threat," *In 55thAIAA/ASMe/ASCE/AHS/SC Structures, Structural Dynamics and Materials Conference*, 2014. pp. 0355.

[4] T. Erlandsson, and L. Niklasson, "A five states survivability model for missions with ground-to-air threats," *In Modeling and Simulation for Defense Systems and Applications VIII*, 2013. pp. 875207.

[5] V. Roberge, M. Tarbouchi, and G. Labonte, *Comparison of parallel genetic algorithm and particle swarm optimization for real-time uav path planning*, IEEE Transactions on Industrial Informatics **9**(1), 132-141 (2013).

[6] Z. Peng, B. Li, X. Chen, and J. Wu, "Online route planning for uav based on model predictive control and particle swarm optimization algorithm," *Proceedings of the 10th World Congress on Intelligent Control and Automation (IEEE)*, 2012. pp. 397–401.
[7] R. W. Beard, T. W. McLain, M. A. Goodrich, and E. P. Anderson, *Coordinated target assignment and intercept for unmanned air vehicles*, IEEE Trans. on Robotics and Automation **18**(6), 911–922 (2002).
[8] W. Xinzeng, C. Linlin, L. Junshan, and Y. Ning, "Route planning for unmanned aerial vehicle based on threat probability and mission time restriction," *In Second IITA International Conference on Geoscience and Remote Sensing (IEEE)*, 2010. pp. 27-30.
[9] B. Başpınar, and E. Koyuncu. "Survivability Based Optimal Air Combat Mission Planning with Reinforcement Learning," *IEEE Conference on Control Technology and Applications (CCTA)*, 2018. pp. 664-669.
[10] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, Cambridge, MA: MIT press, 1998.
[11] R. N. Lothes, M. B. Szymanski, and R. G. Wiley, *Radar vulnerability to jamming*. Norwood, MA: Artech House, 1990.
[12] B. Baspinar, and E. Koyuncu. "Differential Flatness-based Optimal Air Combat Maneuver Strategy Generation," *AIAA SciTech Forum and Exposition, Modelling and Simulation Technologies*, 2019.
[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv.org*, Aug. 28, 2017. [Online].Available: https://arxiv.org/abs/1707.06347. [Accessed: Oct. 27, 2021].

To Cite This Article: B. Başpinar, *Deep Reinforcement Learning-based Cooperative Survivability Maximization for a UAV Fleet on an Air-to-Ground Mission*, Journal of Aeronautics and Space Technologies **15**(2), 94-107 (2022).

VITAE

Barış Başpınar received the B.S. degree in aeronautical engineering and control & automation engineering, the M.S. degree in aerospace engineering, and Ph.D. degree in aerospace engineering from Istanbul Technical University (ITU) in 2013, 2015, and 2020, respectively. He was a Visiting Researcher with the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA, USA, from 2018 to 2019. He is currently an Assistant Professor with the Department of Aeronautical Engineering, ITU. His current research interests include data-driven ATM network modelling, high-level autonomy in air traffic control systems, resiliency problem in large scale complex systems, conflict detection and separation assurance algorithms, air traffic complexity, network theory, survivability analysis for complex air-to-ground missions, air-to-air combat autonomy, optimization-based control of multi-aircraft systems, reinforcement learning, guidance navigation and control problems in aerial vehicles.