ISSN 2076-1546

УДК 004.39 DOI: 10.46972/2076-1546.2025.28.08

І. В. Пулеко, В. О. Чумакевич, В. Б. Ревенко, Д. Є. Ступак, І. В. Свистунович

ПЛАНУВАННЯ ЗАСТОСУВАННЯ БЕЗПІЛОТНИХ ЛІТАЛЬНИХ АПАРАТІВ РОЗВІДКИ МЕТОДАМИ НАВЧАННЯ З ПІДКРІПЛЕННЯМ ТА ВИКОРИСТАННЯМ ФУНКЦІЇ ЩІЛЬНОСТІ РОЗПОДІЛУ ЦІЛЕЙ

Досвід ведення бойових дій у російсько-українській війні переконливо свідчить про актуалізацію проблеми застосування груп безпілотних літальних апаратів для вирішення багатьох військових завдань, зокрема, планування їх ефективного застосування. У свою чергу, виконання цього завдання залежить від системи управління групою та умов застосування. Для децентралізованих систем управління принциповим моментом є необхідність автоматичного планування місії на борту безпілотних літальних апаратів. Аналіз наукових джерел показав, що більшість запропонованих алгоритмів планування не враховують особливостей ведення військової повітряної розвідки й дорозвідки, рельєфу місцевості та здебільшого орієнтовані на оптимізацію польотів за точками. Водночас для планування пошуку цілей застосовують гребінчастий, розширюваний чи вільний способи, однак вони досить незручні для автоматичного планування маршруту самим безпілотним літальним апаратом, оскільки точка на карті зазвичай не співвідноситься з висотою польоту, миттєвим полем зору бортової апаратури спостереження, масштабом та детальністю знімка. Для підвищення можливостей оптимізації автоматичного планування польоту в статті запропоновано застосовувати функцію щільності розподілу цілей. Це динамічна двовимірна математична модель, яка описує умовну відносну ймовірність знаходження цілей у різних точках простору. Вона створюється та задається на основі даних про місцевість, попередніх спостережень або інтелектуальних оцінок, що відображають розподіл можливих цілей на певній ділянці або для всієї зони розвідки. Крім того, вона дозволяє моделювати простір не як однорідний, а як область із різними ступенями важливості або ж імовірностями знаходження цілей.

У рамках цього дослідження розглянуто варіант застосування функції щільності розподілу цілей у навчанні з підкріпленням для планування місій безпілотних літальних апаратів. Навчання з підкріпленням, як один із видів машинного навчання, полягає в навчанні інтелектуального програмного агента (безпілотного літального апарата) приймати рішення щодо послідовності виконуваних дій з урахуванням взаємодії із середовищем для досягнення максимальної винагороди. Щоб реалізувати його в задачі планування застосування безпілотних літальних апаратів в основу процесу математичного моделювання середовища покладено формування ітеративної функції щільності розподілу цілей, яка після невеликої видозміни і буде визначати функцію винагороди. Дією інтелектуального агента на середовище в цьому разі є орієнтація та переміщення його в просторі й отримання знімка підстильної поверхні. Моделювання орієнтації та переміщення безпілотних літальних апаратів у просторі проводиться за допомогою дуальних кватерніонів.

Ключові слова: *агент; безпілотний літальний апарат; функція щільності розподілу цілей; зона розвідки; навчання з підкріпленням; Q-навчання.*

Постановка проблеми в загальному вигляді. Досвід ведення бойових дій у російсько-українській війні переконливо свідчить про актуалізацію проблеми групового застосування безпілотних літальних апаратів (БпЛА) для вирішення багатьох військових завдань [1–2]. Особливо гостро ця проблема постає для груп малих БпЛА, які вже можуть створюватися на основі наявного парку літальних апаратів. При цьому першочерговим завданням, яке повинно бути вирішене для ефективного застосування групи є його планування, яке, у свою чергу, принципово залежить від системи управління групою та умов місії. Тому тема дослідження, що висвітлює розв’язання задачі планування застосування БпЛА розвідки з урахуванням різних умов, є актуальною.

Аналіз останніх досліджень і публікацій. На сьогоднішній день проводиться багато досліджень щодо планування застосування БпЛА, для розвідки передусім це побудова маршрутів руху. Через практичну цінність великої актуальності набувають більш складні задачі, методологія планування маршрутів для яких враховує такі специфічні вимоги та обмеження, як: вплив погодних умов на дальність польоту БпЛА, заборона руху на певних висотах чи заданими районами, визначення місцеположення та наявність зв’язку з іншими учасниками групи тощо.

У найбільш загальному вигляді підходи та класифікація методів планування розглянуті в [3–4]. Стохастичні евристичні алгоритми запропоновано в роботі [5], наведено їх характеристики, напрями вдосконалення, застосування, переваги та недоліки, однак інші види алгоритмів для формування маршрутів руху БпЛА не висвітлено. У публікації [6] алгоритми планування маршруту руху групи БпЛА поділено на п’ять різновидів: оптимізаційні; планування на основі теорії графів; евристичні; ройового інтелекту; нейронно-мережеві. У статті [7] розглянуто алгоритми ройового інтелекту з таких позицій, як: побудова алгоритмів уникнення зіткнень; розподіл завдань; планування маршруту за точками. Планування траєкторії БпЛА в умовах високої щільності перешкод описано у [8]. Планування місця розташування БпЛА на основі стратегії кругового покриття досліджено в [9]. Досить цікавий метод планування маршруту ведення повітряної розвідки динамічних об’єктів з використанням БпЛА подано в [10].

Аналіз останніх досліджень і публікацій показує, що велика кількість вхідних параметрів ускладнює алгоритми та методології розв’язання поставлених задач, оскільки змушує дослідників або не враховувати другорядні параметри взагалі, або ж брати їх до уваги з використанням спрощених алгоритмів. Так, більша частина запропонованих алгоритмів планування не зважає на особливості ведення військової повітряної розвідки й дорозвідки, рельєф місцевості та здебільшого орієнтована на оптимізацію польотів за точками. Водночас для планування пошуку цілей застосовують гребінчастий, розширюваний чи вільний способи, однак вони досить незручні для автоматичного планування маршруту самим БпЛА, оскільки точка на карті зазвичай не співвідноситься з висотою польоту, миттєвим полем зору бортової апаратури спостереження, масштабом та детальністю знімка. Для підвищення можливостей оптимізації автоматичного планування

польоту в [11] запропоновано застосовувати функцію щільності розподілу цілей (ФЩРЦ), тому вважаємо за доцільне розглянути її застосування для різних стратегій управління.

Формулювання завдання дослідження. Метою публікації є дослідження способів використання ФЩРЦ для планування застосування БпЛА в ході виконання розвідувальних завдань. У рамках цього дослідження обмежимося розглядом методів навчання з підкріпленням для планування місій.

ФЩРЦ – це динамічна двовимірна математична модель, яка описує умовну відносну ймовірність знаходження цілей у різних точках простору. Вона створюється та задається на основі даних про місцевість, попередніх спостережень або інтелектуальних оцінок, що відображають розподіл можливих цілей на певній ділянці або для всієї зони розвідки. ФЩРЦ дозволяє моделювати простір не як однорідний, а як область із різними ступенями важливості або ж імовірностями знаходження цілей [11].

Виклад основного матеріалу. Як зазначалося раніше, завдання планування застосування принципово залежить від системи управління групою БпЛА. Для організації таких систем доцільно використовувати деякі загальні управлінські стратегії, зокрема централізоване, децентралізоване та змішане управління. Реалізація навчання з підкріпленням можлива лише в разі децентралізованого управління, яке, у свою чергу, поділяють на колективне та зграйне [12].

Колективне управління передбачає, що в системі немає командира або централізованого пристрою управління, усі одиниці рівноцінні й кожний член групи самостійно ухвалює рішення, намагаючись зробити максимально можливий внесок у досягнення групової мети, при цьому всі члени обмінюються інформацією про обрані дії один з одним. За рахунок того, що кожний елемент вирішує завдання оптимізації лише для себе, а не намагається покращити дії всієї групи, оптимізація суттєво спрощується, тому рішення може знаходитися швидко, у реальному часі. Але це дуже ускладнює алгоритмізацію й висуває до елементів вимогу «високоінтелектуального рівня», тому що вони повинні чітко розуміти групове завдання й уміти вибирати такі дії, які привдуть до найкращого його виконання з погляду всієї групи.

У разі зграйного управління в системі також немає прямого командира або централізованого пристрою управління, усі одиниці рівноцінні й кожний член групи самостійно ухвалює рішення, намагаючись зробити максимально можливий внесок у досягнення групової мети, однак при цьому між членами групи немає обміну інформацією, тому кожен об’єкт «підлаштовує» свої дії на підставі непрямої інформації, слідкуючи за діями інших.

Дуже важливим фактором, що найбільше впливає на вибір можливої стратегії управління, є ступінь автономності БпЛА, з якою тісно пов’язані його автоматичність та інтелектуальність. Лише автономні інтелектуальні БпЛА здатні реалізовувати децентралізовані стратегії та навчання з підкріпленням.

Навчання з підкріпленням (англ. reinforcement learning – RL) – це один із видів машинного навчання, яке полягає в навчанні інтелектуального програмного агента приймати рішення щодо послідовності виконуваних дій з урахуванням взаємодії із середовищем (англ. environment) для досягнення максимальної винагороди (англ. reward) [13].

Щоб реалізувати навчання з підкріпленням у задачі планування застосування БпЛА необхідно визначити його компоненти [13]. Розглянемо їх.

Множина станів середовища (S) описує поточне відображення середовища функціонування БпЛА. Стан середовища може бути поданий як вектор $s \in S$, де S – простір усіх можливих станів.

Множина дій (A) описує конкретні дії, які БпЛА може виконати. Дія подається як $a \in A(s)$, де $A(s)$ – множина можливих дій у стані s. Множина доступних агентові дій є обмеженою (наприклад, БпЛА не може літати далі, ніж на якусь конкретну відстань).

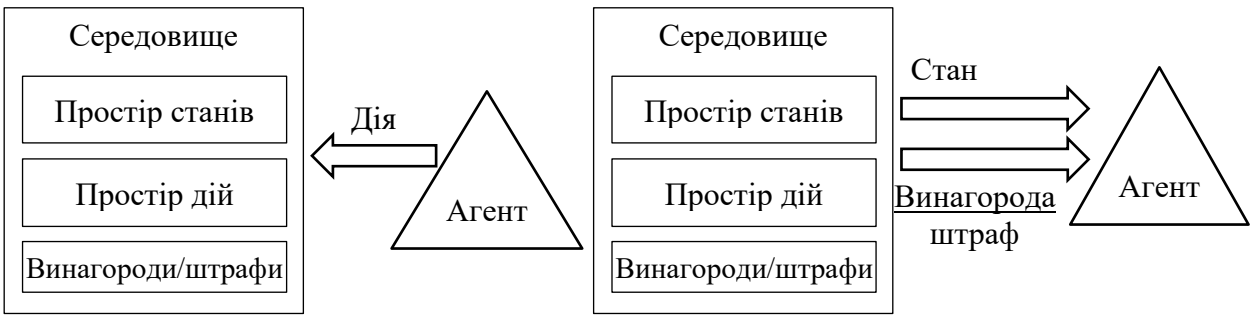
Політика (π) – це правило (або стратегія), за яким БпЛА обирає свої дії, базуючись на поточному стані. Політика $\pi(a|s)$ визначає ймовірність вибору дії a у стані s.

Функція винагороди (R) визначає скалярну винагороду або штраф за виконання певної дії в певному стані. $R(s,a)$ описує винагороду за виконання дії a у стані s.

Правила (P), або ж функція переходу між станами, характеризує ймовірність переходу з одного стану в інший після виконання певної дії. $P(s'|s,a)$ визначає ймовірність переходу в стан s' зі стану s після виконання дії a.

Функція цінності (англ. value – V) оцінює очікувану суму майбутніх винагород, починаючи з певного стану s, слідуючи політиці π. Функція $V^{\pi}(s)$ визначає цінність стану за політики π.

Задача навчання з підкріпленням полягає в навчанні агента (у нашому випадку БпЛА) вибору оптимальних дій у середовищі для максимізації загальної винагороди протягом певного часу. Агент взаємодіє із середовищем (рис. 1а), яке реагує на його дії та змінює свій стан, надаючи агенту винагороду або штраф (рис. 1б) [15].



а) *б)*

Рис. 1. Взаємодія агента та середовища [15]

При цьому для кожної політики розраховується функція цінності:

$$V^{\pi}(s) = \sum_{a \in A(s)} \pi(a|s) \times \sum_{s' \in S} P(s'|s,a)\left[R(s,a) + \gamma V^{\pi}(s')\right], \quad (1)$$

де γ – коефіцієнт дисконтування, що визначає знецінення майбутніх винагород (зазвичай від 0 до 1).

Мета агента – знайти оптимальну політику, яка максимізує очікувану суму винагород:

$$\pi^* = arg\,\max_{\pi} \mathbf{E}\left[\sum_{t=0}^{\infty}\left(\gamma^t R_t|\pi\right)\right], \tag{2}$$

де R_t – винагорода, отримана за час t.

Для оптимізації дій агента в різних станах використовується Q-функція, що визначає очікувану суму винагород за вибір дії в кожному стані та слідування політиці після цього:

$$Q^{\pi}(s,a) = E\left[\sum_{k=0}^{\infty}\left(\gamma^k R_{t+k+1}|s_t = s, a_t = a, \pi\right)\right]. \tag{3}$$

Рівняння для оптимізації Q-функції матиме такий вигляд:

$$Q^{\pi}(s,a) = R(s,a) + \gamma\sum_{s'\in S} P(s'|s,a)\max_{\pi} Q^{\pi}(s',a'), \tag{4}$$

де $P(s'|s,a)$ – імовірність переходу в стан s' після виконання дії a у стані s.

З погляду теорії навчання з підкріпленням складна модель середовища не потрібна. Наприклад, у разі застосування Q-навчання (англ. Q-learning) алгоритм не потребує моделі, його метою є навчитися стратегії, яка вказує агентові, до якої дії вдаватися та за яких обставин. Воно не вимагає моделі середовища (оскільки є безмодельним) і може розв'язувати задачі зі стохастичними переходами та винагородами, не вимагаючи складних обчислень.

Для будь-якого скінченного марковського процесу прийняття рішень (СМППР, англ. finite Markov decision process – FMDP) Q-навчання знаходить стратегію, яка є оптимальною в тому сенсі, що вона максимізує очікуване значення повної винагороди над будь-якими та усіма послідовними кроками, починаючи з поточного стану. Q-навчання може визначати оптимальну стратегію обирання дій для довільного марковського процесу за умови нескінченного часу на розвідування та частково випадкової стратегії. Символом Q позначають функцію, яка повертає винагороду, що використовують для забезпечення підкріплення, і про яку можливо сказати, що вона відповідає «якості» (англ. Quality) дії, обраної в поточному стані [13].

Однак у моделюванні розвідувальних процесів виникає певна суперечність, суть якої полягає в тому, що метою військової розвідки є виявлення, розпізнавання та ідентифікація цілей, тому функція винагороди (або відповідь середовища на дію агента) повинна залежати від кількості виявлених ворожих цілей. Водночас практична реалізація процесів виявлення, розпізнавання та ідентифікації цілей на борту БпЛА досить проблематична. В ідеальних умовах, коли є повне підключення апарата до мережі Інтернет і гарно навчені моделі на основі глибоких штучних нейронних мереж (наприклад, CNN), то в принципі реалізувати виявлення та розпізнавання цілей із певною кількістю помилок цілком можливо. Але в умовах ведення бойових дій та в разі протидії засобів радіоелектронної боротьби поки що це зробити не вдається. Як правило, БпЛА проводить знімання місцевості, а процеси виявлення, розпізнавання та ідентифікації цілей покладені на оператора.

Крім того, оскільки Q-навчання є ітеративним алгоритмом, то воно неявно передбачає якісь початкові умови ще до того, як відбудеться перше уточнення. Постає питання, де їх взяти. Високі початкові цінності, відомі також як «оптимістичні початкові умови» [13], можуть спонукати до дослідження (англ. exploration): не важливо, яку дію обрано, правило уточнення зумовить те, що вона матиме нижчі цінності, ніж інші альтернативи, підвищуючи таким чином імовірність їхнього обрання. Для скидання початкових умов доцільно застосовувати першу винагороду. Відповідно до цієї ідеї в разі першого виконання дії цю винагороду використовують для встановлення значення Q. Це робить можливим негайне навчання за фіксованих детерміністичних винагород.

Тому в основу процесу математичного моделювання середовища, у якому діють БпЛА, пропонуємо покласти формування ітеративної ФЩРЦ [11], яка за невеликої видозміни і буде визначати функцію винагороди. Формування ФЩРЦ починається з карти місцевості. Для отримання моделі в двомірному просторі виділяють мінімальні та максимальні значення по осях координат x і y для всієї області, де визначені дані. Далі встановлюємо крок дискретизації сітки Δx і Δy за кожною з координат. Зону розвідки розбиваємо на сітку з відповідним кроком дискретизації, для кожної ділянки якої обраховуємо значення функції щільності, яке вказує на ймовірність знаходження цілі (цілей) у певній точці. Вхідні дані для побудови такої функції можуть включати інформацію інших видів розвідки, супутникових знімків, історичні відомості про появу цілей у певних зонах, а також моделі поведінки противника. На другому етапі визначаємо зони найвищого інтересу. На основі ФЩРЦ окреслюємо зони з високою ймовірністю появи цілей, тобто зони інтересу. Детально процес побудови ФЩРЦ описано в [11]. У ході функціонування системи розвідки здійснюватиметься доповнення середовища на основі даних, отриманих під час функціонування БпЛА, шляхом внесення відповідних змін до ФЩРЦ.

Отже, у процесі навчання та адаптації БпЛА буде використовуватися динамічне середовище, яке безперервно оновлюється, що дозволить наближати результати моделювання до реальних умов та поліпшувати стратегії взаємодії.

Дією інтелектуального агента (БпЛА) на середовище для задачі, що розглядається, буде орієнтація і переміщення його в просторі та отримання знімка підстильної поверхні. Моделювання орієнтації та переміщення БпЛА в просторі будемо проводити за допомогою дуальних кватерніонів [14].

Орієнтацію БпЛА в просторі визначатимемо за допомогою так званих літакових кутів: рискання ψ, тангажа ϑ і крену γ, – які задають у базовій і зв'язній системі координат.

За базову систему координат приймемо таку: початок системи координат (точка O_0) розташований у точці початку руху БпЛА; вісь O_0X_g спрямована на північ по дотичній місцевого меридіана; вісь O_0Y_g спрямована вертикально вгору і протилежна до напрямку вектора сили тяжіння; вісь O_0Z_g доповнює систему до правої та спрямована праворуч, у бік сходу.

Зв'язна система координат стосується безпосередньо БпЛА: її початок (точка O) розташований у точці центра мас БпЛА; вісь OX спрямована вперед, до передньої точки

БпЛА; вісь OY – вертикально вгору, вона перпендикулярна горизонтальній площині об'єкта; вісь OZ доповнює систему до правої.

Тоді положення БпЛА в просторі задається радіусом-вектором початку (точка *O*) зв'язної системи координат відносно нерухомої базової системи координат. Орієнтація першої відносно другої визначається трьома послідовними поворотами на: кут рискання ψ – поворот навколо осі OY, кут тангажа ϑ – поворот навколо осі OZ, кут крену γ – поворот навколо осі OX.

Для початкового визначення дуального кватерніона необхідно встановити його дійсну й уявну частини. Орієнтація і стан об'єкта задається відносно базової системи координат за допомогою кутів орієнтації ψ, ϑ, γ і вектора положення центра мас $r=\left(r_x,r_y,r_z\right)^T$.

Дійсну частину можна подати за допомогою формули

$$\boldsymbol{q}_1=\begin{bmatrix} cos\frac{\psi}{2}cos\frac{\vartheta}{2}cos\frac{\gamma}{2}-sin\frac{\psi}{2}sin\frac{\vartheta}{2}sin\frac{\gamma}{2} \\ cos\frac{\psi}{2}cos\frac{\vartheta}{2}sin\frac{\gamma}{2}+sin\frac{\psi}{2}sin\frac{\vartheta}{2}cos\frac{\gamma}{2} \\ cos\frac{\psi}{2}sin\frac{\vartheta}{2}sin\frac{\gamma}{2}+sin\frac{\psi}{2}cos\frac{\vartheta}{2}cos\frac{\gamma}{2} \\ cos\frac{\psi}{2}sin\frac{\vartheta}{2}cos\frac{\gamma}{2}-sin\frac{\psi}{2}cos\frac{\vartheta}{2}sin\frac{\gamma}{2} \end{bmatrix}. \tag{5}$$

Потрібно звернути увагу, якщо послідовність поворотів інша, то вирази будуть теж іншими.

Дуальну частину визначаємо за таким виразом:

$$q_2=\frac{1}{2}\,r\otimes q_1. \tag{6}$$

Обчислити кути орієнтації можна з дійсної частини дуального кватерніона q_1:

$$\psi=arctan\frac{2\left(q_0q_2-q_1q_3\right)}{q_0^2+q_1^2-q_2^2-q_3^2},\ \vartheta=arcsin\left(2\left(q_1q_2+q_0q_3\right)\right),\ \gamma=arctan\frac{2\left(q_0q_1-q_2q_3\right)}{q_0^2-q_1^2+q_2^2-q_3^2}. \tag{7}$$

А положення БпЛА обрахуємо в такий спосіб:

$$r=2q_2\otimes q_1^{-1}. \tag{8}$$

У результаті отримуємо вектор у кватерніонній формі $r=\left(0,r_x,r_y,r_z\right)^T$.

Задамо поворот і переміщення вектора дуальним кватерніоном. Для введених систем координат БпЛА ($O_0X_gY_gZ_g$ – нерухомої базової та $OXYZ$ – зв'язної) орієнтацію і його положення відносно базової системи координат можна задати дуальним кватерніоном

$\tilde{q}$ [14]. Якщо заданий вектор r у зв'язній системі координат, то можна отримати вектор r_0 у базовій системі координат за допомогою формули

$$r_0 = \tilde{q} \otimes r \otimes \tilde{q}^{-1}, \tag{9}$$

а також у зворотному напрямку:

$$r = \tilde{q}^{-1} \otimes r_0 \otimes \tilde{q}, \tag{10}$$

де $r = \left(1,0,0,0,0,r_x,r_y,r_z\right)$ – вектор у бікватерніонній формі.

Вибір дії залежить від цілей БпЛА, його поточного стану та умов середовища (ФЩРЦ). Цей підхід дає змогу формалізувати й регулювати поведінку БпЛА в середовищі, забезпечуючи ефективне планування траєкторій руху та адаптивність до змін. Застосування дуальних кватерніонів для опису дій дозволяє точно моделювати поведінку БпЛА і робить можливим використання складних алгоритмів навчання та оптимізації.

Політика інтелектуального агента є набором правил або стратегій, які визначають дії БпЛА залежно від стану середовища та динаміки зміни ФЩРЦ. Вона дозволяє оптимізувати певні задачі, наприклад, максимізацію площі огляду або мінімізацію часу на виконання завдань. Політика може бути описана як функція π, що відображає стан середовища S через дії A, $\pi : S \to A$.

ФЩРЦ відображає стан середовища (координати місцеположення БпЛА), яке постійно змінюється та формує винагороди, що вимагає від БпЛА адаптації своєї стратегії до змін, а саме адекватної реакції на них та прогнозування змін ФЩРЦ. Тому використання навчання з підкріпленням на основі політики агента дає змогу БпЛА ефективно адаптуватися до динамічних обставин, забезпечує гнучкість та високу реактивність системи.

Практичне моделювання процесу навчання проводилося з використанням середовища OpenAI Gym [15]. Процес моделювання включав встановлення необхідних бібліотек Python.

OpenAI Gym пропонує різні ігрові середовища, які можна вбудовувати в наш код та тестувати агента. Бібліотека забезпечує API для надання всієї інформації, потрібної інтелектуальному агенту: можливі дії, розрахунок та поточний стан. Користувачу просто потрібно зосередитися на алгоритмічній частині для свого агента.

На жаль, спеціалізованого середовища для дослідження БпЛА як агента в OpenAI Gym немає. Для моделювання було використано середовище OpenAI Gym під назвою Taxi-V2 [15, 16], звідки було взято всі основні компоненти. Як дії моделювалися рухи за ФШРЦ (північ, південь, захід, схід). За початкову винагороду бралося значення її комірки. Щоб обмежити непотрібні рухи, у такі комірки вписувалося від'ємне значення ФЩРЦ. Процес описує функція

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left(R(s,a) + \gamma \max_{\pi} Q(s',a')\right), \tag{11}$$

де α – це швидкість навчання ($0 < \alpha \leq 1$), тобто ступінь оновлення наших Q-значень на кожній ітерації;

γ – коефіцієнт дисконтування ($0 \leq \gamma \leq 1$), що визначає, наскільки важливими ми хочемо вважати майбутні винагороди. Високе значення коефіцієнта дисконтування (близьке до 1) відображає довгострокову ефективну винагороду, тоді як коефіцієнт дисконтування 0 змушує агента враховувати лише негайну винагороду, що робить його жадібним.

Знак (←) означає оновлення Q-значення поточного стану та дії агента шляхом попереднього визначення ваги $(1-\alpha)$ старого Q-значення, а потім додавання оціненого (вивченого) значення. Вивчене значення – це комбінація винагороди за виконання поточної дії в поточному стані та дисконтованої максимальної винагороди з наступного стану, у якому ми будемо після виконання поточної дії.

Власне, ми вивчаємо правильні дії в поточному стані, оцінюючи винагороду за комбінацію поточного стану / дії та максимальну винагороду для наступного. Це зрештою змусить БпЛА розглядати маршрут з найкращими винагородами разом.

Q-значення пари стан / дія – це сума миттєвої винагороди та дисконтованої майбутньої винагороди (результуючого стану). Для зберігання Q-значення для кожного стану та дії використовуємо Q-таблицю.

Отже, було реалізовано такий алгоритм процесу Q-навчання:

1) ініціалізувати Q-таблицю всіма нулями;

2) почати досліджувати дії: для кожного стану вибрати будь-яку з усіх можливих для поточного стану (S);

3) перехід до наступного стану (s') в результаті дії (a), отримати винагороду з ФЩРЦ;

4) для всіх можливих дій зі стану (s') вибрати ту, у якої найвище значення Q;

5) оновити значення Q-таблиці за допомогою рівняння (11);

6) встановити наступний стан як поточний;

7) якщо досягнуто цільового стану, то завершити процес і повторити його.

Після достатньої кількості випадкового дослідження дій Q-значення мають тенденцію збігатися, слугуючи нашому агенту функцією значення дії, яку він може використовувати для вибору найоптимальнішої із заданого стану.

Є компроміс між дослідженням (вибором випадкової дії) та експлуатацією (вибором дій на основі вже вивчених Q-значень). Ми хочемо запобігти тому, щоб БпЛА завжди рухався одним і тим самим маршрутом, а міг перенавчатися, тому ввели ще один параметр ε, щоб врахувати це під час навчання. Замість того, щоб просто вибирати найкращу вивчену дію з Q-значенням, іноді потрібно надавати перевагу подальшому дослідженню простору дій.

Дослідження показали, що менше значення ε призводить до епізодів із більшою кількістю штрафів (у середньому), що очевидно, оскільки ми досліджуємо та приймаємо випадкові рішення.

Продуктивність агента значно покращилася після Q-навчання, однак, його проблема полягає в тому, що важко реалізувати велику кількість станів у середовищі за допомогою

Q-таблиці, оскільки її розмір стає занадто великим. Вже для розмірів ФЩРЦ 30 × 30 для отримання збіжності необхідно збільшувати кількість перерахунків епізодів приблизно до 100000. І хоча для сучасних обчислювальних засобів це нескладно, усе ж затримки під час навчання доводиться враховувати.

Крім того, у ході одночасного навчання декількох агентів (побудова маршруту для декількох БпЛА) кількість обчислень пропорційно збільшується, тому потрібно додатково формувати алгоритми взаємодії.

Висновки. Двовимірна ФЩРЦ може бути пристосована та застосована для реалізації методів навчання з підкріпленням у задачах побудови маршрутів БпЛА. Такий підхід дає змогу створити адаптивне та гнучке середовище, що відображає реальні умови, він адаптований до місцевості та забезпечує навчання й побудову маршрутів БпЛА.

Важливим моментом є здатність моделі на основі ФЩРЦ динамічно адаптуватися на основі зібраних даних, що дозволяє забезпечити актуальність і релевантність моделі середовища.

Реалізований приклад Q-навчання в цілому дозволяє здійснювати оптимізацію планування маршрутів БпЛА, але не є перспективним для планування маршрутів їх великих груп, оскільки розмір Q-таблиці стає занадто великим.

Перспективним напрямом RL є глибоке навчання з підкріпленням (англ. Deep RL) [17], що передбачає інтеграцію глибоких нейронних мереж для покращення здатності RL-моделей розпізнавати складні шаблони та адаптуватися до складних середовищ. Нейронна мережа приймає інформацію про стан та дії на вхідний шар і навчається виробляти правильні дії з часом.

СПИСОК БІБЛІОГРАФІЧНИХ ПОСИЛАНЬ

1. Застосування безпілотних систем у Силах оборони України : Доктрина ОП 3-0(46). Київ : ЦУБС ГШ ЗСУ, 2023. 56 с.

2. Коршець О., Горбенко В. Уроки застосування безпілотних літальних апаратів у російсько-українській війні // Повітряна міць України. 2023. № 1 (4). С. 9–17. https://doi.org/10.33099/2786–7714–2023–1-4–9-17

3. Чумак О. О., Дудко М. В., Дмітрієв О. М. Онтологія методів планування маршрутів руху безпілотних літальних апаратів // Випробування та сертифікація. 2024. № 1 (3). С. 69–77. https://doi.org/10.37701/ts.03.2024.10

4. Кучеренко О. І., Вакалюк Т. А. Класифікація методів та алгоритмів побудови маршрутів БпЛА // Вчені записки ТНУ ім. В. І. Вернадського. Серія: Технічні науки. 2024. Т. 35 (74), № 3. С. 111–116. https://doi.org/10.32782/2663-5941/2024.3.1/18

5. Stochastic Heuristic Algorithms for Multi-UAV Cooperative Path Planning / Zhao Ch., Liu Yu., Yu L., Li W. // In Proceedings of the 2021 40th Chinese Control Conference (CCC). Shanghai, China, 26–28 July 2021. P. 7677–7682. https://doi.org/10.23919/ccc52363.2021.9549984

6. Sun W., Hao M. A Survey of Cooperative Path Planning for Multiple UAVs // In Proceedings of the International Conference on Autonomous Unmanned Systems. Shanghai, China, 26–28 July 2021. P. 189–196. https://doi.org/10.1007/978-981-16-9492-9_20

7. Tang J., Duan H., Lao S. Swarm Intelligence Algorithms for Multiple Unmanned Aerial Vehicles Collaboration : A Comprehensive Review. Artif. Intell. Rev. 2022. Vol. 56, 4295–4327. https://doi.org/10.1007/s10462-022-10281-7
8. Algorithm for UAV Path Planning in High Obstacle Density Environments: RFA-Star / Zhang W., Li J., Yu W. et al. // Front. Plant Sci. 2024. 15:1391628. https://doi.org/10.3389/fpls.2024.1391628
9. UAV-BS Site Planning Based on Circular Coverage Strategy / Zhang J., Tang Z., Liu X. et al. // Appl. Sci. 2025. 15:1971. https://doi.org/10.3390/app15041971
10. Метод планування маршруту ведення повітряної розвідки динамічних об'єктів з використанням безпілотних літальних апаратів в лісостеповій місцевості / О. І. Тимочко, А. В. Тристан, О. Є. Чернавіна, А. О. Бережний // Системи обробки інформації. 2020. № 3 (162). С. 95–110. https://doi.org/10.30748/soi.2020.162.10
11. Функція щільності розподілу цілей для планування застосування безпілотних літальних апаратів / І. В. Пулеко, В. О. Чумакевич, І. М. Шестак та ін. // Проблеми створення, випробування, застосування та експлуатації складних інформаційних систем : зб. наук. праць. Житомир : ЖВІ, 2024. Вип. 27 (І). С. 69–82. https://doi.org/10.46972/2076-1546.2024.27.06
12. Пулеко І. В. Проблеми управління угрупованням малих безпілотних літальних апаратів з позицій теорії робототехнічних систем // Проблеми створення, випробування, застосування та експлуатації складних інформаційних систем : зб. наук. праць. Житомир : ЖВІ ДУТ, 2015. Вип. 11. С. 106–114.
13. Richard S. Sutton, Andrew G. Barto. Reinforcement Learning : An Introduction. 2nd ed. Cambridge, MA : The MIT Press, 2018. 526 p. ISBN 9780262039246.
14. Модель руху безпілотних літальних апаратів на основі алгебри дуальних кватерніонів / І. В. Пулеко, О. В. Андреєв, О. Ф. Дубина та ін. // Проблеми створення, випробування, застосування та експлуатації складних інформаційних систем : зб. наук. праць. Житомир : ЖВІ, 2023. Вип. 23. С. 52–61. https://doi.org/10.46972/2076-1546.2022.23.04
15. Satwik Kansal, Brendan Martin. Reinforcement Q-Learning from Scratch in Python with OpenAI Gym. URL: https://www.learndatasci.com/tutorials/reinforcement-q-learning-scratch-python-openai-gym/ (last accessed: 18.05.2025).
16. Q-Learning Introduction and Q Table – Reinforcement Learning w/ Python Tutorial p. 1. URL: https://pythonprogramming.net/q-learning-reinforcement-learning-python-tutorial// (last accessed: 18.05.2025).
17. Fan X., Li H., Chen Y., Dong D. UAV Swarm Search Path Planning Method Based on Probability of Containment // Drones. 2024. Vol. 8. P. 132. https://doi.org/10.3390/drones8040132

Стаття надійшла до редакції 19.05.2025.

REFERENCES

1. *Zastosuvannia bezpilotnykh system u Sylakh oborony Ukrainy: Doktryna OP 3-0(46) [Use of Unmanned Systems in the Defense Forces of Ukraine: Doctrine OP 3-0(46)]*. (2023). Kyiv [in Ukrainian].

2. Korshets, O., Horbenko, V. (2023). Uroky zastosuvannia bezpilotnykh litalnykh aparativ u rosiisko-ukrainskii viini [Lessons from the Use of Unmanned Aerial Vehicles in the russo-Ukrainian War]. *Povitriana mits Ukrainy [Air Power of Ukraine], 1* (4), 9–17. https://doi.org/10.33099/2786–7714–2023–1-4–9-17 [in Ukrainian].
3. Chumak, O. O., Dudko, M. V., & Dmitriiev, O. M. (2024). Ontolohiia metodiv planuvannia marshrutiv rukhu bezpilotnykh litalnykh aparativ [Ontology of Route Planning Methods for Unmanned Aerial Vehicles]. *Vyprobuvannia ta sertyfikatsiia [Testing and Certification], 1* (3), 69–77 [in Ukrainian]. https://doi.org/10.37701/ts.03.2024.10
4. Kucherenko, O. I., Vakaliuk, T. A. (2024). Klasyfikatsiia metodiv ta alhorytmiv pobudovy marshrutiv BpLA [Classification of Methods and Algorithms for UAV Route Construction]. *Vcheni zapysky TNU im. V. I. Vernadskoho. Seriia: Tekhnichni nauky [Scientific notes of V. I. Vernadsky TNU. Series: Technical science],* 35 (74), 3, 111–116. https://doi.org/10.32782/2663-5941/2024.3.1/18 [in Ukrainian].
5. Zhao, Ch., Liu, Yu., Yu, L., & Li, W. (2021). Stochastic Heuristic Algorithms for Multi-UAV Cooperative Path Planning. In *Proceedings of the 2021 40th Chinese Control Conference (CCC).* Shanghai, China, July 26–28, 2021. (pp. 7677–7682). https://doi.org/10.23919/ccc52363.2021.9549984
6. Sun, W., & Hao, M. (2021). A Survey of Cooperative Path Planning for Multiple UAVs. In *Proceedings of the International Conference on Autonomous Unmanned Systems.* Shanghai, China, July 26–28, 2021. (pp. 189–196). https://doi.org/10.1007/978-981-16-9492-9_20
7. Tang, J., Duan, H., & Lao, S. (2022). Swarm Intelligence Algorithms for Multiple Unmanned Aerial Vehicles Collaboration : A Comprehensive Review. *Artif. Intell. Rev, 56,* 4295–4327. https://doi.org/10.1007/s10462-022-10281-7
8. Zhang, W., Li, J., & Yu, W. et al. (2024). Algorithm for UAV Path Planning in High Obstacle Density Environments: RFA-Star. *Front. Plant Sci.,* 15:1391628. https://doi.org/10.3389/fpls.2024.1391628
9. Zhang, J., Tang, Z., & Liu, X. et al. (2025). UAV-BS Site Planning Based on Circular Coverage Strategy. *Appl. Sci.,* 15:1971. https://doi.org/10.3390/app15041971
10. Tymochko, O. I., Trystan, A. V., Chernavina, O. Ye., & Berezhnyi, A. O. (2020). Metod planuvannia marshrutu vedennia povitrianoi rozvidky dynamichnykh ob'iektiv z vykorystanniam bezpilotnykh litalnykh aparativ v lisostepovii mistsevosti [The Method of Planning the Route of Conducting Aerial Reconnaissance of Dynamic Objects Using Unmanned Aerial Vehicles in the Forest-Steppe Area]. *Systemy obrobky informatsii [Information Processing Systems], 3* (162), 95–110. https://doi.org/10.30748/soi.2020.162.10 [in Ukrainian].
11. Puleko, I. V., Chumakevych, V. O., & Shestak, I. M. et al. (2024). Funktsiia shchilnosti rozpodilu tsilei dlia planuvannia zastosuvannia bezpilotnykh litalnykh aparativ [Target Density Function For Uav Planning]. *Problemy stvorennia, vyprobuvannia, zastosuvannia ta ekspluatatsii skladnykh informatsiinykh system: zb. nauk. prats [Problems of Construction, Testing, Application and Operation of Complex Information Systems. Scientific Journal of Korolov Zhytomyr Military Institute], 27* (I), 69–82. Zhytomyr: ZhMI. https://doi.org/10.46972/2076-1546.2024.27.06 [in Ukrainian].

12. Puleko, I. V. (2015). Problemy upravlinnia uhrupovanniam malykh bezpilotnykh litalnykh aparativ z pozytsii teorii robototekhnichnykh system [Problems of Group Control by Small Unmanned Aerial Vehicles on Theory Robotic Systems]. *Problemy stvorennia, vyprobuvannia, zastosuvannia ta ekspluatatsii skladnykh informatsiinykh system: zb. nauk. prats [Problems of Construction, Testing, Application and Operation of Complex Information Systems. Scientific Journal of Korolov Zhytomyr Military Institute]*, 11, 106–114. Zhytomyr: ZhMI [in Ukrainian].
13. Richard S. Sutton, & Andrew G. Barto. (2018). *Reinforcement Learning : An Introduction.* 2nd ed. Cambridge, MA: The MIT Press. ISBN 9780262039246.
14. Puleko, I. V., Andreiev, O. V., & Dubyna, O. F. et al. (2023). Model rukhu bezpilotnykh litalnykh aparativ na osnovi alhebry dualnykh kvaternioniv [Model of Motion of Unmanned Aerial Vehicles Based on Dual Quaternion Algebra]. *Problemy stvorennia, vyprobuvannia, zastosuvannia ta ekspluatatsii skladnykh informatsiinykh system: zb. nauk. prats [Problems of Construction, Testing, Application and Operation of Complex Information Systems. Scientific Journal of Korolov Zhytomyr Military Institute], 23,* 52–61. https://doi.org/10.46972/2076-1546.2022.23.04 [in Ukrainian].
15. Satwik Kansal, & Brendan Martin. (n.d.). *Reinforcement Q-Learning from Scratch in Python with OpenAI Gym.* Retrived from https://www.learndatasci.com/tutorials/reinforcement-q-learning-scratch-python-openai-gym/
16. *Q-Learning Introduction and Q Table – Reinforcement Learning w/ Python Tutorial p. 1.* (n.d.). Retrived from https://pythonprogramming.net/q-learning-reinforcement-learning-python-tutorial//
17. Fan, X., Li, H., Chen, Y., & Dong, D. (2024). UAV Swarm Search Path Planning Method Based on Probability of Containment. *Drones, 8,* 132. https://doi.org/10.3390/drones8040132

I. V. Puleko, V. O. Chumakevych, V. B. Revenko, D. Y. Stupak, I. V. Svystunovich

PLANNING THE DEPLOYMENT OF RECONNAISSANCE UNMANNED AERIAL VEHICLES USING REINFORCEMENT LEARNING METHODS AND TARGET DISTRIBUTION DENSITY FUNCTION

The experience of combat operations in the Russo-Ukrainian war clearly highlights the growing relevance of employing groups of unmanned aerial vehicles (UAVs) to accomplish a wide range of military tasks. The primary problem that must be solved to ensure their effective deployment is mission planning. In turn, the planning process depends on the UAV group's control system and the operational conditions. For decentralized control systems, a critical requirement is the capability of onboard autonomous mission planning. A review of the literature indicates that most existing planning algorithms do not account for the specific features of military aerial reconnaissance and target validation, the terrain, and are mostly focused on point-to-point flight optimization. Meanwhile, typical target search planning uses comb, expandable, or free search strategies. However, these approaches are often unsuitable for autonomous route planning by the UAV itself, as coordinates on the map are typically not associated with flight altitude, the instantaneous field of view of onboard sensors, the scale, or image resolution.

To enhance the optimization capabilities of autonomous flight planning, this work proposes the use of a Target Distribution Density Function (TDDF). The TDDF is a dynamic two-dimensional mathematical model that describes the conditional relative probability of target presence at various spatial locations. It is created based on terrain data, prior observations, or intelligent assessments that reflect the distribution of potential targets within a particular area or the entire reconnaissance zone. The TDDF enables modeling the environment not as homogeneous, but as a space with varying levels of importance or likelihood of target presence.

This study explores the use of the TDDF in reinforcement learning for UAV mission planning. Reinforcement learning, a branch of machine learning, involves training an intelligent software agent (the UAV) to make decisions about the sequence of actions by interacting with the environment to maximize cumulative rewards. In the context of UAV mission planning, the environment modeling process is based on the iterative generation of the TDDF, which – with minor adjustments – serves as the reward function.

Keywords: *agent; unmanned aerial vehicle; target distribution density function; reconnaissance zone; reinforcement learning; Q-learning.*