Отримано: 27.12.2025. Доопрацьовано: 02.02.2026. Прийнято: 11.02.2026. Опубліковано: 27.02.2026

УДК 629.735.05(045) DOI 10.31548/energiya1(83).2026.101

БАГАТОАГЕНТНЕ НАВЧАННЯ З ГЛИБОКИМ ПІДКРІПЛЕННЯМ У ЗАДАЧІ ПЛАНУВАННЯ ШЛЯХУ РОЮ БПЛА

В. М. Синєглазов, доктор технічних наук, професор
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»
https://orcid.org/0000-0002-3297-9060
E-mail: svm@kai.edu.ua
І. О. Юденко, аспірант кафедри авіаційних комп'ютерно-інтегрованих комплексів
Національний університет «Київський авіаційний інститут»
https://orcid.org/0000-0002-7022-366X
E-mail: ioyudenko@gmail.com

Анотація. *Робота присвячена багатоагентному навчанню з глибоким підкріпленням у задачі планування шляху. Обгрунтовано використання роїв БПЛА в точному землеробстві. Показано, що для використання роїв дронів необхідно застосовувати штучний інтелект, зокрема навчання з підкріпленням. Поставлена задача планування шляху при наявності поганої якості або вітсутності GPS навігації. Запропоновано використання методу Multi-Agent Proximal Policy Optimization. Отримані результати показали високу якість планування шляху в умовах наявності перешкод та поганої якості або вітсутності GPS навігації.*

Ключовіслова: ***безпілотні літальні апарати, точне землеробство, мультиагентні системи, штучний інтелект***

Вступ. Нині в наслідок війни в Україні скоротилася площа посівних земель на 20-25 %. Також варто відмітити, що велика частка пахотних земель залишається замінованою і по світовим даним, для її розмінування потрібно буде принаймні 100 років.

Тому для забезпечення підняття продуктивності залишку пахотних земель необхідно значно підняти їх продуктивність. Єдиним варіантом вирішення даної проблеми є точне землеробство. Зрештою, точне землеробство зменшує трудовитрати, підвищує операційну ефективність, запобігає втратам урожаю та зміцнює рентабельність рослинництва та тваринництва.[1]

Вирішення задач точного землеробства можливе завдяки дронам, які поєднують доступність, маневреність та високоточну навігацію з RGB-, мультиспектральними і тепловими камерами, а також системами автоматизованої обробки даних. Таке оснащення дозволяє швидко змінювати зони інтересу, працювати на малій висоті, отримувати детальні знімки, недосяжні супутникам і пілотованій авіації, оперативно картографувати поля, виявляти стресові зони посадок, фіксувати осередки захворювань рослин, оцінювати потреби у зрошенні та добривах і контролювати структуру ґрунтового покриву, включно з ранніми проявами ерозії.

Однак використання одиночних дронів недостатньо ефективно Тому суцільно використовувати рої дронів які беруть на себе функції синхронного обстеження угідь, точкового внесення агрохімікатів та постійного контролю за станом посівів.Рійбезпілотних літальних апаратів (БПЛА) - це група з безлічі БПЛА, часто невеликих і недорогих, які працюють спільно та автономно (або напівавтономно) для досягнення спільної мети.

Серед базових технічних завдань роєвого підходу особливе місце займає планування траєкторії: огляд застосування ШІ до цього завдання [2] та роботи з планування руху роїв БПЛА [3] демонструють, як алгоритми будують безпечні та енергоефективні маршрути у динамічних умовах. Дослідження з координаційного управління [4] пропонує інтегроване планування, що охоплює весь рій, а проблема локалізації вирішується за допомогою високорівневих схем, заснованих на оцінці положення RSSI [5].

Однак рої дронів це складні для ручного управління динамічні системи, тому слід використовувати ШІ.

Штучний інтелект кардинально змінює перспективи БПЛА, дозволяючи їм виконувати завдання з високим ступенем автономії та ефективності. В найближчому майбутньому ШІ посилюватиме здібності БПЛА у виявленні та відстежуванні цілей, плануванні траєкторій та автономної навігації, включаючи польоти у складних або закритих просторах. Розвиток глибоких нейромереж (CNN, RNN) та методів посиленого навчання розширить можливості для спостереження, пошуково-рятувальних операцій та точного сільськогосподарського моніторингу за рахунок покращеного аналізу зображень та відео в реальному часі. Координація роїв за допомогою інтелектуальних алгоритмів відкриє сценарії колективних місій, що масштабуються, – від інспекцій і картографування до оперативних військових і рятувальних дій.

Огляд літературних джерел.БПЛА мають скритність, мобільність і низьку вартість, тому сфери їх застосування постійно розширюються, а умови польоту стають складнішими. У зв'язку з цим автономне планування траєкторії перетворюється на критично важливий елемент системи прийняття рішень, що визначає безпеку та ефективність роботи БПЛА.

Зі зростанням використання БПЛА у завданнях стеження, порятунку при НС та моніторингу навколишнього середовища зростає роль алгоритмів планування траєкторії [6,7]. Однак реальні, динамічні умови польоту висувають високі вимоги до багатоблочних систем, і помилки планування можуть призводити до серйозних збитків.

Класичні підходи (штучне потенційне поле [8,9], алгоритм A* [10], генетичний алгоритм [11], MPC [12]) будуються за схемою: формування карти – планування траєкторії – управління рухом. Через високу трудомісткість побудови карти такі методи погано підходять для завдань навігації у реальному часі. Їм притаманні складність реалізації, низька продуктивність у реальному часі, схильність до локальних мінімумів і слабка адаптивність до середовища, що швидко змінюється.

Альтернативою є методи навчання з підкріпленням (RL), які розглядають навігацію як завдання послідовного ухвалення рішень у рамках марківської моделі. Алгоритм, взаємодіючи із середовищем, знаходить оптимальну функцію цінності «стан-дія» і, тим самим, оптимальну послідовність дій. Для БПЛА це дозволяє виконувати реактивну навігацію без попередньої побудови карти з динамічним уникненням перешкод та плануванням у реальному часі.

У цій роботі вирішується завдання планування траєкторії при невідомому місцезнаходження мети та відсутності GPS. Запропоновано алгоритм, що поєднує візуальну навігацію та навчання з підкріпленням для групи БПЛА.

По-перше, оскільки точні координати мети недоступні, використовується алгоритм YOLOV5 [13] для її виявлення та оцінки положення зображення. Потім об'єднані візуальні ознаки та вихідне зображення подаються на вхід RL-алгоритму планування траєкторії.

По-друге, застосовується розподілений підхід: кожен БПЛА сприймається як незалежний агент зі своєю політикою і приймає рішення лише з урахуванням власних спостережень, без межаппаратной зв'язку.

По-третє, розроблено імітаційну платформу та проведено апаратно-орієнтовані експерименти. Результати показують, що запропонований метод забезпечує успіх, який можна порівняти з випадком, коли точні координати мети відомі.

Останніми роками дедалі більше досліджень присвячено алгоритмам планування шляху з урахуванням навчання з підкріпленням [14]. Навчання з підкріпленням пов'язує стани (State) з діями (Action) та спрямоване на максимізацію винагороди (Reward) у заданому середовищі: агент спостерігає середовище, вибирає дію з набору допустимих, отримує винагороду та прагне виробити оптимальну політику. Схеманавчаннязпідкріпленнямпоказанорисунку1.

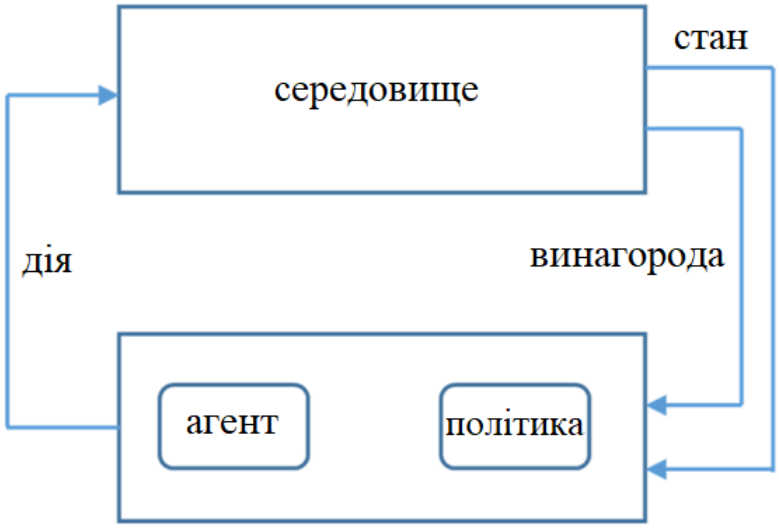

Рис. 1. Концепція навчання із підкріпленням

Планування шляху природно формалізується як завдання послідовного прийняття рішень та відповідає марківському процесу прийняття рішень (МПД). МПД задається четвіркою (S, A, P, R), де S – простір станів, A – безліч дій, P – розподіл переходів, R – функція винагороди, а політика π відображає стан у дію. Дисконтована сумарна винагорода для цієї політики визначається стандартною формулою з коефіцієнтом. На її основі вводяться функція цінності Vπ(s) та функція цінності стану-дії Qπ(s,a); при відомому розподілі переходів $\boldsymbol{p(s_t+1|s_t,a_t)}$ та винагороді $\boldsymbol{r(s_t,a_t)}$ з рівняння Беллмана виводяться відповідні співвідношення. Мета агента - наблизити оптимальну функцію Q*(s,a) і на її основі отримати оптимальну політику. Навчання з підкріпленням відбувається через цикл «проба – помилка – коригування» і показало свою ефективність для завдань уникнення перешкод та планування траєкторій: достатньо навчання на вибірках середовища для отримання працездатної моделі.

З аналізу випливає, більшість алгоритмів планування траєкторії припускають відоме точне становище мети і формують функцію винагороди з його основі. У разі відсутності GPS таке припущення недоступне, що унеможливлює застосування багатьох методів. Хоча є роботи, що враховують відсутність GPS, вони вирішують завдання для одного БПЛА і не розглядають взаємодію кількох апаратів за відсутності зв'язку. Потрібно розробити алгоритми планування траєкторій для багатовимірного угруповання БПЛА за умов відсутності GPS та зв'язку.

Мета дослідження - розробити метод багатоагентного планування шляху рою БПЛА під час виконання агрокультурних робіт землеробства, що передбачає планування шляху при наявності перешкод та поганого або відсутнього зв’язку.

Матеріали та методи дослідження. Багатоагентне навчання з підкріпленням (MARL) переносить ідеї підкріплювального навчання (RL) на системи з кількома автономними агентами, які взаємодіють між собою та з одним спільним оточенням. У класичному RL агент через спроби й помилки вчиться обирати дії в певних станах, щоб максимізувати накопичену винагороду; MARL застосовує цей підхід до випадків із двома й більше агентами, які можуть співпрацювати, конкурувати або поєднувати обидва режими, прагнучи оптимізувати індивідуальні або спільні цілі.

Потреба в MARL зумовлена тим, що багато реальних складних систем мають багатоагентну структуру – наприклад, рої дронів, роботизовані команди або автономні транспортні засоби. MARL критично важливе для налагодження координації, спільної роботи та конкурентної взаємодії в середовищах, де агенти суттєво впливають одне на одного.

MARL створює низку специфічних проблем, які не властиві одноагентним RL-середовищам:

- Нестаціонарність: для кожного агента середовище змінюється, бо політики інших агентів теж еволюціонують.
- Призначення кредиту: важко встановити, які дії конкретного агента призвели до загального результату.
- Масштабованість: простори станів і дій зростають експоненційно з кількістю агентів, ускладнюючи обчислення.
- Часткова спостережуваність: агенти часто мають неповну інформацію про глобальний стан або внутрішні стани інших.

Алгоритми MARL зазвичай поділяються натригрупи: методи, що базуються на значеннях (value-методи), методи, що базуються на політиці (policy-методи)табагатоагентніметодиглибокогонавчаннязпідкріпленням (MADRL).

Багатоагентні методи DRL розширюють перераховані парадигми на середовища з декількома агентами, що взаємодіють. Приклади включають QMIX, MADDPG, MATD3 та MAPPO. Вони вирішують завдання децентралізованого прийняття рішень та нестаціонарної динаміки, застосовуючи техніки накшталт централізованого навчання з децентралізованим виконанням, факторизації функції цінності та спільного використання параметрів.

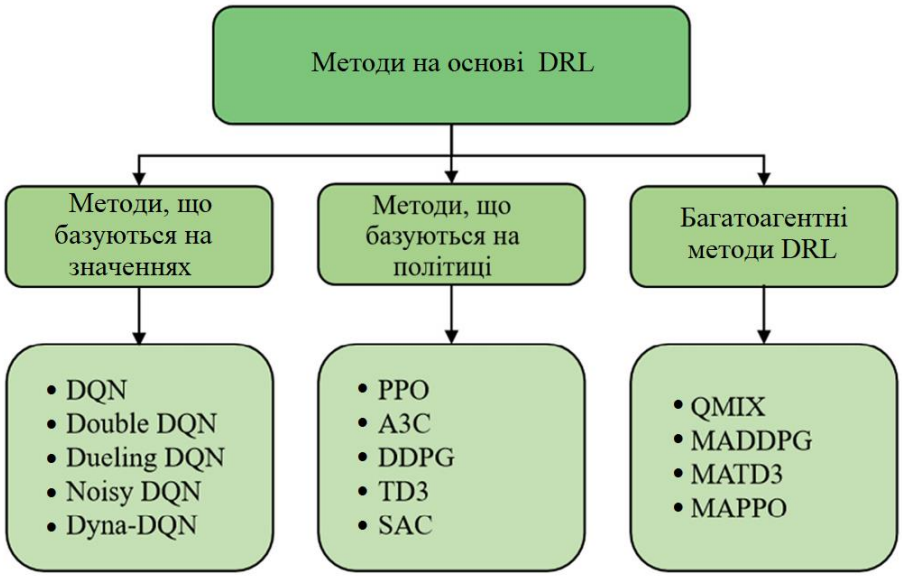

Рис. 2. Класифікація методів запобігання зіткнень на основі глибокого навчання із підкріпленням

Багатоагентні методи глибокого навчання з підкріпленням (MADRL)

MADRL переносить одноагентні підходи в розраховані на багато користувачів середовища, де потрібно враховувати нестаціонарність, часткову спостерігальність і координацію. Серед характерних алгоритмів:

(1) QMIX

QMIX - кооперативний value-факторизаційний алгоритм, який представляє глобальну функцію цінності через монотонну нелінійну комбінацію локальних Q-функцій агентів. Така факторизація дозволяє навчати централізовано, але виконувати децентралізовано, зберігаючи узгодження локальних та глобальних оптимумів.

(2) MADDPG

MADDPG розширює DDPG для багатоагентного середовища, використовуючи централізованих критиків, які мають доступ до спільних спостережень та дій на етапі навчання. Кожен агент має власний актор для децентралізованого виконання, а критик навчається з урахуванням інформації всіх агентів, що підвищує стійкість і координацію за умов нестаціонарності.

(3) MATD3

MATD3 – багатоагентна версія TD3, що вводить подвійних критиків та відкладені оновлення політик для зниження зміщення переоцінки та покращення збіжності. Кожен агент має актора і пару централізованих критиків; до цільових дій додається усічений Гаусів шум для згладжування цільової політики, що полегшує навчання критика в безперервному дійсному просторі.

(4) MAPPO

MAPPO адаптує PPO під багатоагентний сценарій, застосовуючи CTDE та загальні структури критиків.Усічена сурогатна цільова функція забезпечує стабільні оновлення політики, а спільне використання параметрів критика при збереженні індивідуальних політик забезпечує баланс між координацією та масштабованістю.

Визначення проблеми

1. БПЛА повинні триматися подалі від оборонної зони під час польоту; вхід у неї призведе до їх знищення або захоплення.

2. Оборонна зона створюється електромагнітними засобами та невидима для ока.

3. Якщо GPS недоступний, БПЛА знає лише загальну площу цілі, а не її точні координати.

4. Якщо зв'язок заборонено, БПЛА не можуть обмінюватися даними та можуть лише визначати положення та стан інших БПЛА за допомогою бортових камер.

Поширені методи виявлення об'єктів поділяються на дві основні групи: двоетапні підходи, прикладами яких є сімейство RCNN, та одноетапні підходи, типовими для сімейства YOLO. Виявлення об'єктів шукає всі області інтересу на зображенні та виводить їх розташування та ймовірності класів. YOLO розглядає виявлення як задачу регресії: одна згорткова нейронна мережа обробляє все зображення, розділяє його на сітку та прогнозує обмежувальні рамки та ймовірності класів для кожної комірки. YOLO є швидким, оскільки він усуває складний багатоетапний конвеєр, вирішуючи виявлення за один прохід.

Наразі YOLOv5 є найсучаснішим у сфері точності та швидкості виявлення. Ми обрали YOLOv5 для розпізнавання зображень під час планування траєкторії руху кількох БПЛА. Його магістраль включає модулі Focus та C3Net, причому два різних варіанти C3Net використовуються для магістралі та головки детектування. Функція втрат YOLOv5 наведена нижче.

$$loss = \lambda_1 L_{cls} + \lambda_2 L_{obj} + \lambda_3 L_{loc} \quad (1)$$

$L_{cls}, L_{obj}, L_{loc}$ позначають втрату категорії, втрату позитивної та негативної вибірки та втрату локації відповідно. $\lambda_1, \lambda_2, \lambda_3$ позначають коефіцієнти рівноваги. У цій роботі ми вибірково вибрали цільову позицію, позначили її, а потім навчили алгоритм YOLOv5 для отримання ваги виявлення.

Однак, безпосереднє використання обмежувальної рамки виявлення створює серйозну проблему часткової спостережуваності. На початку місії БПЛА може не виявити ціль, створюючи постійні нульові спостереження та не надаючи корисної інформації для дослідження чи планування шляху. Щоб вирішити цю проблему, ми об'єднуємо дані зображення та ознаки виявлення як вхідні дані та застосовуємо багатоагентний підхід PPO під назвою MAPPO. Навчання з підкріпленням отримує як необроблене зображення, щоб БПЛА міг сприймати зону протиповітряної оборони та інші БПЛА, так і вихідні дані розпізнавання цілі, щоб вказати місцезнаходження цілі. Ці два вхідні дані об'єднуються та обробляються спільно, покращуючи ситуаційну обізнаність кожного БПЛА та зменшуючи часткову спостережуваність. Потім об'єднані вхідні дані подаються в алгоритм MAPPO для створення політик та функцій стан-значення для агентів.

Схема MAPPO представлена на рисунку 3 [15]. У MAPPO кожен агент *i* навчає децентралізовану політику $\pi\theta_i(a_i|s_i)$, де θ_i - параметри дії a_i на основі свого локального спостереження s_i, при цьому під час навчання використовується централізований критик $V_\phi(s)$, що має доступ до глобального стану **s**.

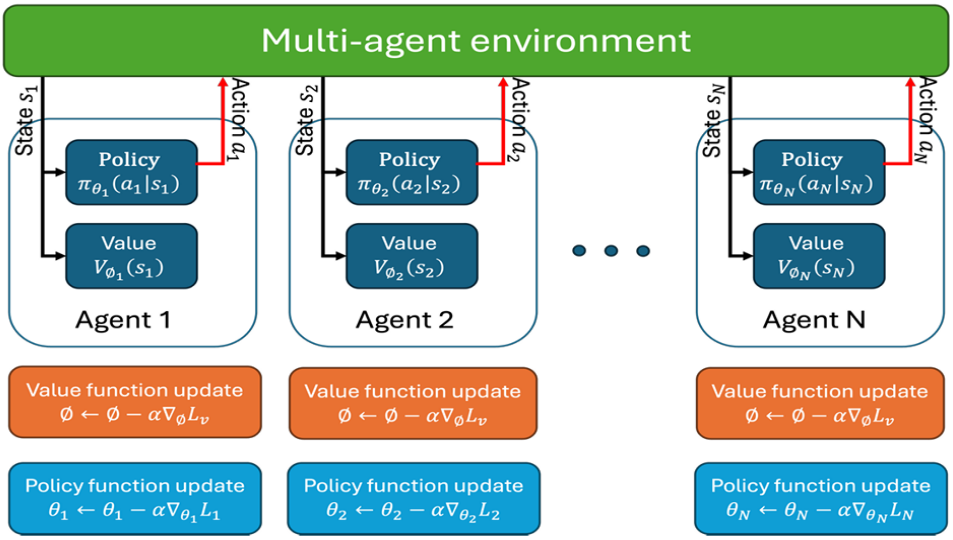

Рис.3. Багатоагентна оптимізація політики на близькій відстані (MAPPO)

Алгоритм MAPPO (багатоагентна проксимальна оптимізація політик) має наступний вигляд

Initialize policy parameters $\theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ for n agents
Initialize centralized value function parameters ϕ
for each iteration **do**
  for each environment rollout **do**
    for $t = 1$ to T **do**
      for each agent $i = 1$ to N **do**
        Observe local state s_i
        Sample action $a_i \sim \pi_{\theta_i}(a_i|s_i)$
      end for
      Execute joint action $a = (a_1, \ldots, a_n)$
      Observe next state s and joint reward $r = (r_1, \ldots, r_N)$
      Store (s, a, r, s') in buffer
    end for
  end for
  for each agent $i = 1$ to N **do**
    Compute advantages $\hat{A}_i$ using Generalized Advantage Estimation (GAE):

$$\delta_i = r_i + \gamma V_\phi(s') - V_\phi(s)$$

$$\hat{A}_i = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_i^{t+l}$$

    Compute the PPO objective:

$$L_i = \min\left(\frac{\pi_{\theta_i}(a_i|s_i)}{\pi_{\theta_i}^{\text{old}}(a_i|s_i)}\hat{A}_t^i, \text{clip}\left(\frac{\pi_{\theta_i}(a_i|s_i)}{\pi_{\theta_i}^{\text{old}}(a_i|s_i)}, 1-\epsilon, 1+\epsilon\right)\hat{A}_i\right)$$

    Update policy parameters:

$$\theta_i \leftarrow \theta_i + \alpha\nabla_{\theta_i} L_i$$

  end for
  Update value function parameters by minimizing:

$$L_v = \sum_t (V_\phi(s) - R)^2$$

$$\phi \leftarrow \phi - \alpha\nabla_\phi L_v$$

end for

Мета навчання кожного агента адаптована з функції втрат з обмеженням (clippedsurrogateloss) алгоритму PPO, яка стабілізує оновлення політики, перешкоджаючи суттєвим відхиленням між новою та попередньою політикою. Функція втрат для агента і задається так:

$$L_i(\theta_i) = \mathbb{E}\left[\min\left(r_i(\theta_i)\hat{A}_i, clip(r_i(\theta_i), 1-\epsilon, 1+\epsilon)\hat{A}_i\right)\right], \quad (2)$$

де відношення ймовірностей $r_i(\theta_i)$ задається формулою

$$r_i(\theta_i) = \frac{\pi_{\theta_i}(a_i|s_i)}{\pi_{\theta_i}^{old}(a_i|s_i)}, \quad (3)$$

а $\hat{A}_i$ означає функцію переваги для агента і, яка оцінюється з використанням узагальненої оцінки переваги (GAE):

$$\delta_i = r_i + \gamma V_\phi(s') - V_\phi(s), \quad (4)$$

$$\hat{A}_i = \sum_{l=0}^{T-t}(\gamma\lambda)^l \delta_i^{t+l}. \quad (5)$$

Тут γ - коефіцієнт дисконтування, r_i – винагорода, δ – це заздалегідь визначена межа KL-дивергенції, а λ регулює компроміс між зміщенням та дисперсією при оцінці переваг.

Централізована функція цінності $V_\phi(s)$ оновлюється шляхом мінімізації середньоквадратичної помилки між прогнозованим значенням та фактичною прибутковістю R:

$$L_v = \sum_t \left(V_\phi(s) - R\right)^2, \quad (6)$$

Результати досліджень та їх обговорення. У таблиці представлені критерії для багатоагентного навчання з підкріпленням (MARL) з акцентом на керування БПЛА, показуючи, як потреби середовища, застосування та продуктивності впливають на доцільнісит вибору методу. MARL на основі політик (наприклад, MAPPO) безпосередньо оптимізує політики та підходить для завдань безперервного керування, таких як польоти в групі та кооперативні маневри. Цей підхід є гнучкими та надійними у складних розподілених сценаріях (периферний штучний інтелект, рої БПЛА) та добре справляється з обмеженнями конфіденційності та зв'язку, але зазвичай потребує більшої кількості епізодів навчання та може повільніше сходитися через асинхронні оновлення та затримки зв'язку.Час навчання сильно залежить від кількості агентів (спільне зростання простору дій), типу алгоритму (методи політики жертвують довшою конвергенцією заради стабільності), накладних витрат на зв'язок (централізовані/гібридні системи), складності середовища (динамічні перешкоди, часткова спостережуваність) та структури винагород (розріджена проти щільної). Тому різні алгоритми пропонують додаткові переваги щодо масштабованості, конвергенції, ефективності вибірки та обмежень розгортання. MAPPO, зокрема, демонструє сильну масштабованість та конвергенцію для керування групою з узагальненою оцінкою переваг, проте остаточний вибір методу повинен визначатися специфічними для місії факторами, такими як відмовостійкість, готовність до розгортання та обмеження обчислень.

Критерії для методів навчання MARL на основі політики

Критерій	**MARL наоснові політик**
Основна ідея	Безпосередньо вивчає децентралізовані або централізовані політики
Репрезентативні алгоритми	MAPPO, MADDPG, TRPO, HAPPO
Простір дій	Безперервний та дискретний
Продуктивність	Високий рівень у безперервних, динамічних завданнях
Середовище	Добре працює в динамічних, частково спостережуваних або безперервних середовищах
Масштабованість	Високий (децентралізоване виконання ефективне)
Витрати на зв'язок	Помірний (залежить від спільного використання політик)
Стабільність навчання	Більш стабільний за умови належного дослідження (наприклад, GAE)
Швидкість конвергенції	Помірний; вимагає більше вибірок, але стабільний
Адаптивність до динамічних середовищ	Високий
Збереження конфіденційності	Не підтримується за своєю природою
Придатність для застосування БПЛА	Ідеально підходить для безперервного контролю (наприклад, формування, відстеження)

Висновки.У роботі обгрунтовано необхідність використання роїв БПЛА в точному землеробстві. На основі проведеного аналізу запропоновано вкористання машинного навчання з підкріпленням для розв'язання задачі планування маршруту рою в складних умовах, а саме відсутності GPS та наявності перешкод (ліси, складний рельєф).

Список використаних джерел

1. IoT-Enabled Precision Agriculture: Developing an Ecosystem for Optimized Crop Management. / S. Atalla, et al. *Information*. 2023; 14(4):205. https://doi.org/10.3390/info14040205

2. A review of artificial intelligence applied to path planning in UAVswarms. / A. Puente-Castro, et al. Neural Comput. Appl. 2022, 34, 153–170.

3. Motion Planning of UAV Swarm: Recent Challenges and Approaches. In Aeronautics-New Advances / M. M. Iqbal et al. IntechOpen: London, UK, 2022 .

4. Zhu, X.; Liu, Z.; Yang, J. Model of collaborative UAV swarm toward coordination and control mechanisms study. Procedia Comput.Sci. 2015, 51, 493–502.

5. Paulsson, M. High-Level Control of UAV Swarms with RSSI Based Position Estimation. Master's Thesis, Lund University, Lund,Sweden, 2017.

6. Toward Autonomous UAV Swarm Navigation: A Review of Trajectory Design Paradigms. / K. Arshid et al. Sensors, 25(18), 5877. https://doi.org/10.3390/s25185877.

7. Poudel, S.; Moh, S. Task assignment algorithms for unmanned aerial vehicle networks: A comprehensive survey. Veh. Commun. 2022, 35, 100469.

8. Khatib, O. Real-time bstacle avoidance for manipulators and mobile robots. Int. J. Robot. Res. 1986, 5, 90–98.

9. An improved artificial potential field method for path planning and formation control of the multi-UAV systems. / Z. Pan et al. IEEE Trans. Circuits Syst. II Express Briefs 2022, 69, 1129–1133.

10.Self-optimization A-star algorithm for UAV path planning based on Laguerre diagram. / R. Wei et al.Syst. Eng. Electron 2015, 37, 577–582.

11.Holland, J.H. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence; MIT Press: Cambridge, MA, USA, 1975.

12.Zhenhua, P.; Hongbin, D.; Li, D. A multilayer graph for multi-agent formation and trajectory tracking control based on MPC algorithm. IEEE Trans. Cybern. 2021, 50, 12.

13.Yolov5. 2022. Available online: https://github.com/ultralytics/yolov5

14.Sutton, R.S.; Barto, A.G. Reinforcement Learning, 2nd ed.; An Introduction; MIT Press: Cambridge, MA, USA, 2018.

15.Jiang, J.; Contributors, M. MARLlib Documentation: PPO Family of Algorithms. 2023. Available at: https://marllib.readthedocs.io/en/latest/algorithm/ppo_family.html

MULTI-AGENT DEEP REINFORCEMENT LEARNING IN THE PATH PLANNING PROBLEM

V. Sineglazov, I. Yudenko

Abstract. *This work is devoted to multi-agent deep reinforcement learning in the path planning problem. The use of UAV swarms in precision agriculture is justified. It is shown that for the use of drone swarms it is necessary to apply artificial intelligence, in particular reinforcement learning. The task of path planning in the presence of poor quality or absence of GPS navigation is set. The use of the Multi-Agent Proximal Policy Optimization method is proposed. The results obtained showed high quality of path planning in the presence of obstacles and poor quality or absence of GPS navigation.*

Keywords: ***unmanned aerial vehicles, precision agriculture, multi-agent systems, artificial intelligence***

References

1. Atalla S, Tarapiah S, Gawanmeh A, Daradkeh M, Mukhtar H, Himeur Y, Mansoor W, Hashim KFB, Daadoo M. (2023). IoT-Enabled Precision Agriculture: Developing an Ecosystem for Optimized Crop Management. *Information*, 14(4):205. https://doi.org/10.3390/info14040205

2. Puente-Castro, A.; Rivero, D.; Pazos, A.; Fernandez-Blanco, E. (2022). A review of artificial intelligence applied to path planning in UAVswarms. Neural Comput. Appl. 34, 153–170. [CrossRef]

3. Iqbal, M.M.; Ali, Z.A.; Khan, R.; Shafiq, M. (2022). Motion Planning of UAV Swarm: Recent Challenges and Approaches. In Aeronautics-New Advances; IntechOpen: London, UK .

4. Zhu, X.; Liu, Z.; Yang, J. (2015). Model of collaborative UAV swarm toward coordination and control mechanisms study. Procedia Comput.Sci., 51, 493–502. [CrossRef]

5. Paulsson, M. (2017). High-Level Control of UAV Swarms with RSSI Based Position Estimation. Master's Thesis, Lund University, Lund,Sweden.

6. Arshid, K., Krayani, A., Marcenaro, L., Gomez, D. M., & Regazzoni, C. (2025). Toward Autonomous UAV Swarm Navigation: A Review of Trajectory Design Paradigms. Sensors, 25(18), 5877. https://doi.org/10.3390/s25185877.

7. Poudel, S.; Moh, S. (2022). Task assignment algorithms for unmanned aerial vehicle networks: A comprehensive survey. Veh. Commun., 35, 100469.

8. Khatib, O. (1986). Real-time bstacle avoidance for manipulators and mobile robots. Int. J. Robot. Res., 5, 90–98.

9. Pan, Z.; Zhang, C.; Xia, Y.; Xiong, H.; Shao, X. (2022). An improved artificial potential field method for path planning and formation control of the multi-UAV systems. IEEE Trans. Circuits Syst. II Express Briefs, 69, 1129–1133.

10. Wei, R.; Xu, Z.; Wang, S.; Lv, M. (2015). Self-optimization A-star algorithm for UAV path planning based on Laguerre diagram. Syst. Eng. Electron, 37, 577–582.

11. Holland, J.H. (1975). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence; MIT Press: Cambridge, MA, USA

12. Zhenhua, P.; Hongbin, D.; Li, D. (2021). A multilayer graph for multi-agent formation and trajectory tracking control based on MPC algorithm. IEEE Trans. Cybern, 50, 12.

13. Yolov (2022). Available online: https://github.com/ultralytics/yolov5

14. Sutton, R.S.; Barto, A.G.(2018). Reinforcement Learning, 2nd ed.; An Introduction; MIT Press: Cambridge, MA, USA.

15. Jiang, J.; Contributors, M. MARLlib (2023). Documentation: PPO Family of Algorithms. 2023. Available at: https://marllib.readthedocs.io/en/latest/algorithm/ppo_family.html