Simposio CEA de Robótica, Bioingeniería, Visión por Computador y Automática Marina 2026

Marco metodológico para la navegación de robots en invernaderos mediante aprendizaje por refuerzo

Cañadas-Aránega, Fernando[a,*], Gil, Juan D.[a], Moreno-Úbeda, José C.[a], Blanco-Claraco, José L.[b]

[a] *University of Almeria, Department of Informatics, CIESOL, ceia3, Ctra. Sacramento s/n, 04120, Almería, Spain. (fernando.ca, juandiego.gil, jcmoreno)@ual.es*
[b] *University of Almeria, Department of Engineering, CIESOL, ceia3, Ctra. Sacramento s/n, 04120, Almería, Spain. jlblanco@ual.es*

Resumen

La navegación autónoma de robots móviles en invernaderos agrícolas es un reto debido a la naturaleza dinámica, semiestructurada y parcialmente observable de estos entornos. Este artículo propone un marco metodológico basado en el aprendizaje por refuerzo para la planificación de trayectorias en invernaderos mediterráneos. El problema se formula como un planificador sin modelo y se resuelve utilizando el algoritmo *Deep Deterministic Policy Gradient* (DDPG), empleando un esquema de entrenamiento offline basado en el conjunto de datos GREENBOT. Se define una función de recompensa específica para los invernaderos agrícolas, que equilibra el progreso hacia el objetivo con penalizaciones por colisiones y comportamientos inseguros. También se propone un procedimiento de ajuste fino en línea para mejorar la adaptación a las variaciones estructurales y dinámicas del entorno. Este trabajo establece las bases teóricas y metodológicas para el futuro desarrollo de sistemas de navegación robustos en invernaderos reales.

Palabras clave: Robótica agrícola, Machine Learning, Planificación de trayectorias, ROS2 Humble, Navegación autónoma

Methodological framework for robot navigation in greenhouses using reinforcement learning

Abstract

Autonomous navigation of mobile robots in agricultural greenhouses is challenging due to the dynamic, semi-structured, and partially observable nature of these environments. This paper proposes a methodological framework based on reinforcement learning for trajectory planning in Mediterranean greenhouses. The problem is formulated as a model-free planner problem and is solved using the Deep Deterministic Policy Gradient (DDPG) algorithm, employing an offline training scheme based on the GREENBOT dataset. A reward function specific to agricultural environments is defined that balances progress toward the goal with penalties for collisions and unsafe behaviors. An online fine-tuning procedure is also proposed to improve adaptation to structural and dynamic variations in the environment. This work establishes the theoretical and methodological foundations for the future development of robust navigation systems in real greenhouses.

Keywords: Agricultural robotics, Machine learning, Path planning, ROS2 Humble, Automatic navigation

1. Introducción

La agricultura bajo invernadero es uno de los entornos más exigentes para la navegación autónoma de robots móviles. En las últimas décadas, la superficie mundial dedicada a este tipo de cultivo ha superado las 490.000 hectáreas, con un crecimiento sostenido impulsado por la necesidad de aumentar la productividad, garantizar la seguridad alimentaria y reducir la dependencia de mano de obra humana (Trenda, 2023; Moreno et al., 2024). En particular, el modelo de invernadero de tipo Mediterráneo, predominante en regiones como el sureste de España, concentra una parte significativa de esta superficie y se caracteriza por niveles de tecnificación bajos o medios. En este contexto, la incorporación de robots móviles autónomos se perfila como una solución clave para mejorar la eficiencia operativa y la competitividad del sector agrícola (Cañadas-Aránega et al., 2024). A diferencia de los entornos industriales controlados, los invernaderos constituyen espacios semi-estructurados, dinámicos y altamente variables. Aunque la disposición en filas de cul-

*Corresponding author: fernando.ca@ual.es

tivo introduce cierta regularidad geométrica, la navegación del robot se ve afectada por factores como el crecimiento continuo de las plantas, la presencia de hojas y ramas que generan oclusiones, la coexistencia con trabajadores humanos, la circulación de maquinaria agrícola y condiciones de iluminación cambiantes. Estas características convierten la planificación de trayectorias en un problema complejo, donde la incertidumbre, el ruido sensorial y la aparición de obstáculos dinámicos son inherentes al entorno (Cañadas-Aránega et al., 2025).

La navegación autónoma de robots móviles en invernaderos requiere la generación de trayectorias seguras y eficientes que permitan alcanzar objetivos específicos —como puntos de inspección, recolección o tratamiento— evitando colisiones y respetando las restricciones físicas del entorno. Tradicionalmente, este problema se ha abordado mediante métodos clásicos de planificación de trayectorias, basados en mapas previos, modelos geométricos del entorno y funciones de coste diseñadas manualmente. Sin embargo, estos enfoques suelen asumir entornos estáticos o completamente conocidos, así como una localización precisa del robot, condiciones que rara vez se cumplen en escenarios agrícolas reales (Blanco-Claraco et al., 2023; Cañadas-Aránega et al., 2024).

En este escenario, el aprendizaje por refuerzo (RL, del inglés, *Reinforcement Learning*) emerge como una alternativa especialmente adecuada para la planificación de trayectorias en invernaderos. Al formular la navegación como un problema de toma de decisiones secuencial, el agente de RL puede aprender políticas de control directamente a partir de la interacción con el entorno, optimizando su comportamiento en presencia de incertidumbre y cambios dinámicos. Esta capacidad resulta especialmente relevante en invernaderos, donde el entorno no puede considerarse estático ni completamente predecible (Xin et al., 2017).

En este contexto, se han utilizado algoritmos de planificación global basados en mapas previamente construidos, combinados con planificadores locales para la evitación de obstáculos en pasillos estrechos. Los algoritmos de tipo actor-crítico para espacios continuos han adquirido un papel destacado (Gil et al., 2026). En particular, el algoritmo *Deep Deterministic Policy Gradient* (DDPG) ha sido ampliamente utilizado en robótica móvil debido a su capacidad para generar políticas deterministas de problemas de control de forma *offline*, es decir, entrenando al agente DDPG a partir de un conjunto de datos previamente recolectados, sin necesidad de interactuar con el entorno. Esta característica lo convierte en una opción especialmente adecuada para robots agrícolas que deben operar bajo estrictas restricciones cinemáticas (Wang et al., 2023). Sin embargo, la sensibilidad de DDPG a la elección de hiperparámetros y su limitada robustez frente a ruido sensorial han motivado la exploración de alternativas más estables. Como alternativa, *Proximal Policy Optimization* (PPO) ha demostrado una mayor estabilidad durante el entrenamiento gracias a su formulación basada en actualizaciones acotadas de la política, facilitando su aplicación en escenarios complejos como los invernaderos (Wu et al., 2026). Por otro lado, *Soft Actor–Critic* (SAC) introduce un criterio de maximización de entropía que favorece la exploración eficiente y la robustez frente a incertidumbre, características especialmente relevantes en entornos agrícolas dinámicos (Yu et al., 2024). A pesar de los avances logrados con estos algoritmos, su aplicación sistemática a robots móviles en invernaderos sigue siendo limitada, lo que evidencia la necesidad de adaptar y evaluar estas técnicas considerando las restricciones geométricas, dinámicas y sensoriales propias de la agricultura bajo invernadero.

No obstante, la adopción de un agente basado en DDPG permite aprovechar varias ventajas frente a los métodos alternativos mencionados anteriormente. En primer lugar, su estructura actor-crítico con políticas deterministas continuas permite un control de las acciones, evitando la discretización del espacio de control y reduciendo la pérdida de información asociada. Otra ventaja relevante es la posibilidad de entrenar el agente en modo *offline*, empleando datos previamente recolectados, lo que disminuye el riesgo de daños en el hardware durante la fase de aprendizaje y permite aprovechar registros de operación reales provenientes de sistemas expertos.

Este trabajo propone un marco metodológico para la planificación reactiva de trayectorias basada en la política aprendida de robots móviles en invernaderos agrícolas, mediante el algoritmo DDPG. El agente aprenderá una política determinista que transformará observaciones sensoriales en comandos de velocidad lineal y angular, maximizando una función de recompensa diseñada específicamente para equilibrar el avance hacia el objetivo, la evitación de colisiones y la suavidad del movimiento en pasillos estrechos. La principal contribución de este estudio radica en el establecimiento de una formulación matemática y estructural adaptada a entornos agrícolas bajo invernadero, así como en la definición de un procedimiento de entrenamiento fuera de línea basado en datos reales provenientes de un sistema experto, que servirá como base para futuras implementaciones en plataformas robóticas reales.

El artículo tiene la siguiente estructura: la sección 2 describe la configuración experimental que se propondrá en estudios futuros; la sección 3 realiza un análisis del sistema planificador por RL; finalmente, la sección 4 muestra las conclusiones de la propuesta.

2. Configuración experimental

En esta sección se describen todo los modelos de interacción con el entorno.

2.1. Invernadero Agroconnect

Para llevar a cabo las simulaciones presentadas en este estudio, se ha utilizado un modelo tridimensional del invernadero desarrollado en (Cañadas-Aránega et al., 2026), desarrollado a partir de las instalaciones de Agroconnect situadas en el municipio de La Cañada de San Urbano (Almería, España). Este invernadero tiene una extensión de 40x36 m y cuenta con once pasillos, flanqueados a ambos lados por plantas de tomate y separados entre sí 4 m, conformando así los corredores de navegación del robot. El modelo 3D reproduce fielmente un cultivo real de tomate tipo pera en sistema hidropónico, así como la geometría característica de un invernadero mediterráneo, proporcionando un entorno realista para la validación de los algoritmos de navegación.

Como entorno de simulación se ha empleado el MultiVehicle Simulator (MVSim) (Blanco-Claraco et al., 2023) (véase la Fig. 1), ya que incorpora modelos de fricción realistas basados

en física para la interacción neumático–suelo, lo que resulta especialmente adecuado para el análisis de perturbaciones y el desarrollo de leyes de control.

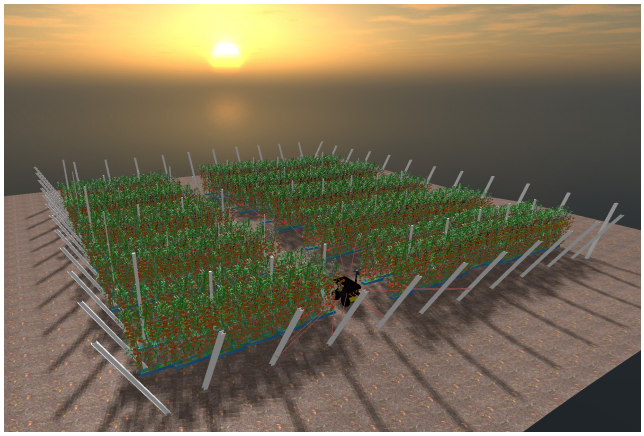

Figura 1: Modelo 3D del invernadero en MVSim

2.2. *Robot AgriCobIoT I*

Uno de los vehículos autónomos que se emplean es el robot móvil diferencial de cuatro ruedas AgriCobIoT I (Moreno Úbeda et al., 2022), basado en el modelo A200 de Clearpath, aunque estructuralmente adaptado para la realización de tareas agrícolas colaborativas en entornos de invernadero. Este sistema constituye la plataforma principal de control, cuya dinámica se describe mediante los estados de un robot diferencial $\mathbf{x}_D(t) = [x_D(t); y_D(t); \theta_D(t),] \in \mathbb{R}^3$, con $t \in \mathbb{R}^+$, definida por la ecuacion:

$$\begin{aligned} \dot{x}_D(t) &= v_D(t)\cos(\theta_D(t)), \\ \dot{y}_D(t) &= v_D(t)\sin(\theta_D(t)), \\ \dot{\theta}_D(t) &= \omega(t), \end{aligned} \tag{1}$$

donde $v_D(t) \in \mathbb{R}$ y $\omega(t) \in \mathbb{R}$ denotan la velocidad lineal longitudinal y la velocidad angular del robot. En la Figura 2 se muestra el modelo 3D implementado.

Figura 2: Robot AgriCobIoT I en MVSim

2.3. *Robot AgriCobIoT II*

El otro robot que se pretende integrar es el robot AgriCobIoT II, un robot terrestre de cuatro ruedas con arquitectura Ackermann, diseñado específicamente para la navegación y realización de tareas agrícolas colaborativas en entornos de invernadero. Bajo el supuesto de rodadura pura y deslizamiento despreciable, la cinemática del robot Ackermann, cuya dinámica se describe mediante el estado $\mathbf{x}_A(t) = [x_A(t); y_A(t); \theta_A(t)] \in \mathbb{R}^3$ definido por la ecuacion:

$$\begin{aligned} \dot{x}_A(t) &= v_A(t)\cos(\theta_A(t)), \\ \dot{y}_A(t) &= v_A(t)\sin(\theta_A(t)), \\ \dot{\theta}_A(t) &= \frac{v_A(t)}{L_A}\tan(\delta(t)), \end{aligned} \tag{2}$$

donde $v_A(t) \in \mathbb{R}$ representa la velocidad lineal longitudinal del vehículo, $\delta(t) \in \mathbb{R}$ es el ángulo de giro de las ruedas delanteras, y $L_A \in \mathbb{R}^+$ denota la distancia entre ejes del robot. Este modelo cinemático resulta especialmente adecuado para describir el comportamiento del robot AgriCobIoT II en desplazamientos suaves y maniobras de navegación dentro de los pasillos del invernadero. En la Figura 3 se muestra el modelo 3D implementado.

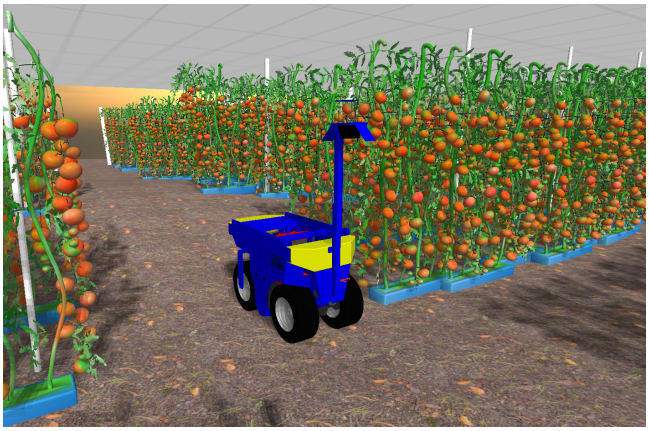

Figura 3: Robot AgriCobIoT II en MVSim

3. Sistema de navegación por RL

A continuación se detallan todos los conceptos y características del sistema RL.

3.1. *Conocimientos previos*

La mayoría de los enfoques de aprendizaje por refuerzo empleados en la navegación de robots móviles en invernaderos modelan el entorno como un Proceso de Decisión de Markov (MDP, del inglés *Markov Decision Process*), descrito mediante la quíntupla $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ (Gil et al., 2026). En esta formulación, $\mathcal{S}$ representa el conjunto de estados del sistema, $\mathcal{A}$ el conjunto de acciones que el robot puede ejecutar, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ la función de transición entre estados, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ la función de recompensa, y γ el factor de descuento que pondera la influencia de recompensas futuras. Sin embargo, en escenarios de operación reales en un invernadero,

el estado completo del sistema rara vez es directamente accesible, debido principalmente a oclusiones causadas por la vegetación, ruido en las mediciones sensoriales y restricciones en el campo de visión de los sensores embarcados. Como consecuencia, el agente de RL toma decisiones a partir de observaciones parciales del entorno, denotadas por el espacio O, lo que conduce a una formulación más realista del problema como un Proceso de Decisión de Markov Parcialmente Observable - POMDP, del inglés *Partially Observable Markov Decision Process*. En este marco, el objetivo del agente es aprender una política óptima basada en la información perceptiva disponible, siguiendo el esquema clásico de interacción agente–entorno para la navegación autónoma en invernaderos agrícolas. En la Figura 4 se muestra el esquema que sigue el RL propuesto.

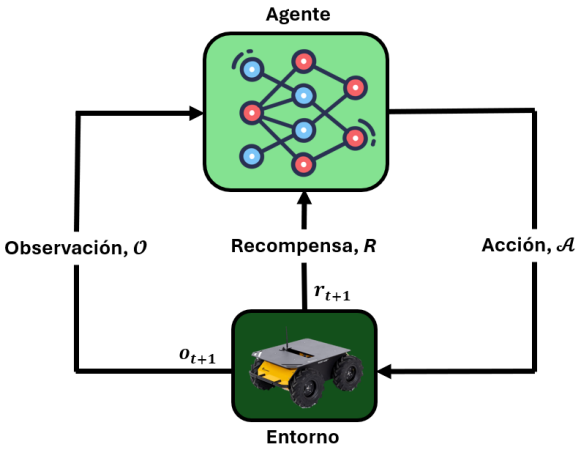

Figura 4: Aprendizaje por refuerzo: el algoritmo RL actualiza la política y el entorno devuelve un estado siguiente y una recompensa por cada acción. El algoritmo RL utiliza la recompensa y el estado siguiente para actualizar la política y seleccionar mejores acciones.

3.2. *Deep Deterministic Policy Gradient*

Entre los algoritmos más representativos que emplean una arquitectura de tipo actores críticos en problemas con espacios de acción continuos destaca el método DDPG (Gil et al., 2026). Este algoritmo se basa en el uso de dos redes neuronales diferenciadas: una red *actor*, encargada de generar acciones deterministas a partir de las observaciones del entorno, y una red crítica, cuya función es estimar cómo de buena es la acción ejecutada mediante una función Q.

Durante el proceso de entrenamiento fuera de línea del agente, cuya definición se realiza en la próxima subsección, las experiencias obtenidas a partir de la interacción con el entorno del sistema experto, se almacenan en una memoria denominada *buffer de experiencias*. Cada experiencia está compuesta por las observaciones del sistema, las acciones ejecutadas y las recompensas recibidas en instantes de tiempo discretos. Posteriormente, el entrenamiento de las redes se realiza extrayendo mini-lotes aleatorios de tamaño M desde dicho *buffer*, lo que permite reducir la correlación temporal entre muestras y mejorar la estabilidad del aprendizaje. La actualización de la red critica se realiza a partir de un mini-lote de transiciones extraídas del *buffer de experiencias*.

Sea $i \in \{1, \ldots, M\}$ el índice que recorre las $M \in \mathbb{N}^+$ muestras seleccionadas en cada iteración de entrenamiento. Para cada transición, el valor objetivo se define como

$$y_i^{(t)} = r_i + \gamma Q_{\Phi'}(o_{i+1}, \pi_{\theta_a^t}(o_{i+1})), \tag{3}$$

donde y_i es el valor objetivo en i, r_i representa la recompensa inmediata asociada a i, $\gamma \in (0, 1)$ es el factor de descuento que pondera la influencia de las recompensas futuras, $o \in O$ representa el conjunto de observaciones, θ_a^t corresponde con los parámetros en el indice de entrenamiento t de la red actor objetivo en redes objetivo (*target networks*), y Φ representa el conjunto de parámetros entrenables. En aplicaciones de navegación móvil, valores típicos de γ se sitúan entre 0.95 y 0.99, lo que promueve comportamientos orientados al largo plazo sin comprometer la estabilidad numérica. Las funciones $Q_{\Phi'}$ y $\pi_{\theta_a^t}$ corresponden a las redes objetivo t del crítico y del actor, respectivamente, cuyos parámetros se actualizan de forma suave mediante interpolación para mejorar la estabilidad del aprendizaje.

A partir de estos valores objetivo, la red crítica se entrena minimizando la siguiente función de pérdida cuadrática media:

$$L(\Phi) = \frac{1}{M} \sum_{i=1}^{M} \left(y_i^{(t)} - Q(o_i, a_i; \Phi) \right)^2, \tag{4}$$

Este procedimiento ajusta la función $Q(o, a; \Phi)$ (con $a \in \mathcal{A}$ representando el conjunto de acciones) para que aproxime el valor esperado acumulado asociado a cada par observación–acción, proporcionando una estimación consistente del desempeño futuro del robot durante la navegación en el entorno de invernadero.

Por su parte, la red actor se actualiza maximizando la recompensa acumulada esperada, utilizando un gradiente aproximado calculado sobre un mini-lote de tamaño M:

$$\nabla_\theta J \approx \frac{1}{M} \sum_{i=1}^{M} G_{a_i} G_{\pi_i}, \tag{5}$$

donde G_{a_i} representa el gradiente de la salida de la red crítica con respecto a la acción generada, y G_{π_i} corresponde al gradiente de la salida de la red actor con respecto a sus propios parámetros. Esta formulación permite adaptar políticas de control del sistema experto de manera eficiente, resultando especialmente adecuada para la navegación de robots móviles en entornos agrícolas bajo invernadero.

3.3. *Offline RL*

La metodología planteada se estructura en tres fases principales, que se describen de forma esquemática en el Algoritmo 1.

Tabla 1: Resumen del *GREENBOT dataset* utilizado para entrenamiento y evaluación.

Sensor	**Tipo de datos**	**Frecuencia / Resolución**
Cámara estéreo	Imágenes RGB (pares)	10 Hz (1032 × 776)
LiDAR Velodyne VLP16	Nube de puntos 3D	10 Hz (360° H FoV)
LiDAR Ouster OS0	Nube de puntos 3D + IMU	10 Hz (275° H FoV)
IMU	Aceleración / Giros	Integrado en LiDAR OS0

Algoritmo 1: Desarrollo del agente DDPG para la navegación de robots móviles en invernaderos con entrenamiento fuera de línea y ajuste fino en tiempo real

Entrada: *Dataset* de información correspondiente al invernadero (Cañadas-Aránega et al., 2024), cuyos datos se detallan en la tabla 1.

Salida: Política de navegación optimizada para el desplazamiento autónomo del robot móvil en entornos de invernadero.

Fase 1: Preparación del *dataset*:

1. Recolectar datos históricos de navegación del robot en invernaderos reales o simulados, incluyendo información de sensores (LiDAR, cámaras de profundidad, odometría) y estados del robot.
2. Preprocesar los datos recopilados y organizarlos en tuplas de transición (o_t, a_t, r_t, o_{t+1}).

Fase 2: Entrenamiento fuera de línea del agente DDPG:

1. Inicializar aleatoriamente los parámetros de las redes *actor* y *critic* del agente DDPG.
2. Inicializar las redes objetivo y el *buffer* de experiencias con el *dataset*.
3. Aplicar el proceso de entrenamiento fuera de línea utilizando mini-lotes extraídos del *buffer*.

Fase 3: Ajuste fino del DDPG:

1. Desplegar la política entrenada en el robot móvil dentro del invernadero.
2. Recolectar nuevas experiencias en línea durante la navegación autónoma, considerando la presencia de obstáculos dinámicos y variaciones sensoriales.
3. Actualizar dinámicamente el *buffer* de experiencias incorporando las nuevas transiciones.
4. Realizar el ajuste fino del agente DDPG mediante actualizaciones periódicas de las redes *actor* y *critic*.

En la Figura 5 se muestra una representación 3D de la trayectoria como odometría del dataset, obtenida como el valor estimado de *Inertial Measurement Unit* (IMU) junto a la nube de puntos, donde se muestra cada una de las colisiones detectadas con puntos rojos.

Como se muestra en el esquema propuesto, el proceso se iniciará utilizando el dataset GREENBOT (Cañadas-Aránega et al., 2024), que contiene experiencias reales de navegación en un invernadero mediterráneo e incluye observaciones sensoriales multimodales y acciones ejecutadas por el robot bajo distintas configuraciones del entorno. Este conjunto de datos servirá como base para el entrenamiento fuera de línea de una política DDPG, permitiendo establecer un comportamiento inicial seguro y coherente con las restricciones del entorno agrícola.

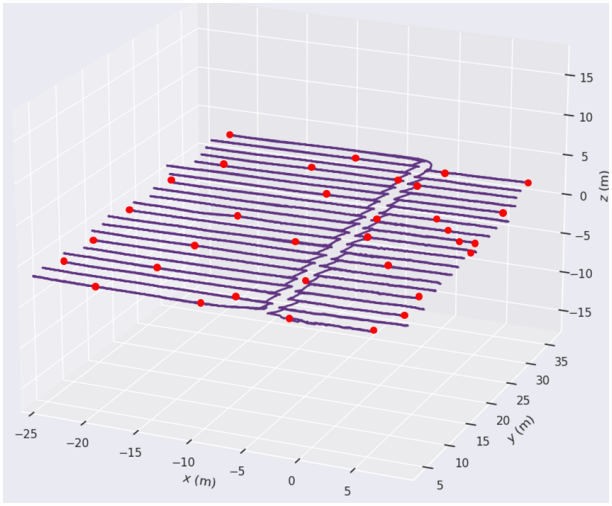

Figura 5: Trayectoria del datashet, cuyos puntos rojos representan las colisiones registradas para el entrenamiento *Offline*.

Posteriormente, se plantea la incorporación de un ajuste fino en línea que permita adaptar la política aprendida a variaciones estructurales, presencia de obstáculos dinámicos y cambios perceptivos no contemplados en el conjunto de datos original. De este modo, el marco propuesto combina aprendizaje basado en datos históricos con capacidad de adaptación futura en condiciones reales de operación.

3.4. *Función de recompensa*

La función de recompensa $r_t \in \mathbb{R}$ propuesta en este trabajo se muestra en la Ecuación (6). Esta formulación ha sido diseñada específicamente para reflejar los objetivos operativos de la navegación en invernaderos agrícolas, incentivando el progreso hacia el objetivo y penalizando situaciones de riesgo registradas en los datos históricos.

En el contexto del entrenamiento fuera de línea planteado, el agente no interactuará inicialmente con el entorno real, sino que aprenderá a partir de transiciones previamente recolectadas. En consecuencia, los eventos de colisión presentes en el conjunto de datos se utilizarán como ejemplos de comportamiento indeseado, guiando el aprendizaje hacia políticas más seguras.

La estructura de recompensa densa basada en la distancia al objetivo permitirá proporcionar señales informativas en cada transición, lo cual resulta especialmente relevante en escenarios agrícolas con percepción parcial y ruido sensorial. Esta formulación será evaluada en trabajos futuros en simulación y validación sobre plataforma real.

$$\text{recompensa } r_t = \begin{cases} r_{\text{llegada}}, & \text{si el robot alcanza el TP,} \\ r_{\text{colisión}} - d_t^2, & \text{colisión en el dataset,} \\ -d_t^2, & \text{en otro caso.} \end{cases} \tag{6}$$

En esta formulación, d_t representa la distancia euclídea entre el robot y el punto objetivo en el instante t. El término r_{llegada} corresponde a una recompensa positiva asociada a la finalización satisfactoria de la tarea, mientras que $r_{\text{colisión}}$ es una penalización negativa aplicada cuando en el conjunto de datos se observa una transición que conduce a colisión. El término continuo $-d_t^2$ actúa como recompensa densa, incentivando la reducción progresiva de la distancia al objetivo. Esta formulación favorece un aprendizaje estable y reduce problemas de extrapolación en aprendizaje fuera de línea, especialmente en entornos agrícolas con percepción parcial y ruido sensorial.

Cabe destacar que el presente trabajo tiene como objetivo establecer el marco conceptual y metodológico para la aplicación de aprendizaje por refuerzo en entornos agrícolas bajo invernadero. La implementación completa, evaluación cuantitativa y validación experimental sobre plataformas reales se abordarán en trabajos futuros, permitiendo analizar métricas como tasa de éxito, número de colisiones, suavidad de trayectoria y capacidad de generalización.

4. Conclusiones

En este trabajo se ha presentado una formulación basada en aprendizaje por refuerzo para la planificación de trayectorias de robots móviles en entornos agrícolas bajo invernadero. A diferencia de los enfoques clásicos de planificación global, que requieren mapas completos y modelos geométricos estáticos, la propuesta modela el problema como un proceso de decisión secuencial parcialmente observable, permitiendo al robot aprender políticas de control directamente a partir de datos sensoriales.

La metodología se ha fundamentado en el algoritmo DDPG, adecuado para espacios de acción donde las salidas del agente corresponden directamente a comandos de velocidad lineal y angular del robot. Se ha definido explícitamente el espacio de observaciones y acciones en relación con la dinámica de plataformas diferenciales y Ackermann, así como una función de recompensa diseñada para equilibrar el avance hacia el objetivo, la seguridad frente a colisiones y la suavidad del movimiento en pasillos estrechos característicos de los invernaderos mediterráneos. La formulación matemática desarrollada proporciona un marco coherente y reproducible para aplicar técnicas de aprendizaje por refuerzo en robótica agrícola.

Actualmente, se está trabajando en la definición completa del espacio de acciones y del espacio de observaciones del sistema, con el fin de finalizar la configuración del proceso de entrenamiento y proceder a la realización de pruebas experimentales en un invernadero real.

Agradecimientos

Este trabajo ha sido financiado por el Proyecto LIFE ACCLIMATE, LIFE23-CCA-ES-LIFE ACCLIMATE, cofinanciado por la Unión Europea bajo el acuerdo de concesión del programa LIFE No. LIFE23-CCA-ES-LIFE-ACCLIMATE/101157315, y ha usado la infraestructura ”AgroConnect.es” (ayuda EQC2019-006658-P) para llevar a cabo esta investigación, financiada por MCIN/AEI/10.13039/501100011033 y por FEDER Una manera de hacer Europa. Además, el autor Fernando Cañadas-Aránega cuenta con una beca FPI (PRE2022-102415) del Ministerio de Ciencia, Innovación y Universidades.

Referencias

Blanco-Claraco, J.-L., Tymchenko, B., Mañas-Alvarez, F. J., Cañadas-Aránega, F., López-Gázquez, Á., Moreno, J. C., 2023. Multivehicle simulator (mvsim): Lightweight dynamics simulator for multiagents and mobile robotics research. SoftwareX 23, 101443.

Cañadas-Aránega, F., Border, R., Blanco-Claraco, J. L., Moreno Úbeda, J. C., 2025. Agriculture surface edge explorer (agrisee): Active reconstruction of occluded tomatoes in greenhouses. Simposios del Comité Español de Automática (CEA) 1 (1).

Cañadas-Aránega, F., Mañas-Álvarez, F. J., Moreno, J. C., Blanco-Claraco, J. L., et al., 2026. A ros2 benchmarking framework for hierarchical control strategies in mobile robots for mediterranean greenhouses. arXiv preprint arXiv:2602.15162.

Cañadas-Aránega, F., Moreno, J. C., Blanco-Claraco, J. L., 2024. A pid-based control architecture for mobile robot path planning in greenhouses. IFAC-PapersOnLine 58 (7), 503–508.

Cañadas-Aránega, F., Blanco-Claraco, J. L., Moreno, J. C., Rodriguez-Diaz, F., 2024. Multimodal mobile robotic dataset for a typical mediterranean greenhouse: The greenbot dataset. Sensors 24 (6).

Gil, J. D., Chanona, E. A. D. R., Guzmán, J. L., Berenguel, M., 2026. Reinforcement learning meets bioprocess control through behavior cloning: Real-world deployment in an industrial photobioreactor. Engineering Applications of Artificial Intelligence 164, 113326.

Moreno, J., Rodríguez, F., Sánchez-Hermosilla, J., Giménez, A., Sánchez-Molina, J., 2024. Feasibility analysis of robots in greenhouses. a case study in european mediterranean countries. Smart Agricultural Technology 9, 100638.

Moreno Úbeda, J. C., Cañadas-Aránega, F., Rodríguez, F., Sánchez-Hermosilla, J., Giménez, A., 2022. Modelado 3d y diseño de un robot colaborativo para tareas de transporte en invernaderos. In: XLIII Jornadas de Automática. Universidade da Coruña. Servizo de Publicacións, pp. 785–791.

Trenda, E., 2023. Greenhouse agricultural area in spain in 2022, by type of crop. `https://www.statista.com/statistics/1218871/greenhouse-area-spain-by-crop/`, accessed: 22 April 2023.

Wang, Y., He, Z., Cao, D., Ma, L., Li, K., Jia, L., Cui, Y., 2023. Coverage path planning for kiwifruit picking robots based on deep reinforcement learning. Computers and Electronics in Agriculture 205, 107593.

Wu, B., Ding, Z., Ostigaard, L., Huang, J., 2026. Reinforcement learning-based energy-aware coverage path planning for precision agriculture. arXiv preprint arXiv:2601.16405.

Xin, J., Zhao, H., Liu, D., Li, M., 2017. Application of deep reinforcement learning in mobile robot path planning. In: IEEE 2017 Chinese Automation Congress (CAC). pp. 7112–7116.

Yu, L., Chen, Z., Wu, H., Xu, Z., Chen, B., 2024. Soft actor-critic combining potential field for global path planning of autonomous mobile robot. IEEE Transactions on Vehicular Technology 74, 7114 – 7123.