CONTROL AUTÓNOMO DE DRON UTILIZANDO COMANDOS DE VOZ Y APRENDIZAJE POR REFUERZO ASISTIDO

Francisco Cruz
Doctor en Ciencias de la Computación

Rubén Contreras
Ingeniero Civil en Computación e Informática

En los últimos años, la tecnología relacionada con vehículos aéreos no tripulados (UAV) ha logrado expandir el conocimiento en esta área, dando paso a nuevos problemas y desafíos que requieren solución. Además, esta tecnología muestra una gran demanda en la industria, permitiendo automatizar procesos que realizan las personas normalmente. El aprendizaje por refuerzo (RL) como marco de automatización, es frecuentemente usado para entrenar agentes autónomos. RL es un paradigma de aprendizaje de máquina donde un agente interactúa con un ambiente para resolver una determinada tarea. Sin embargo, el aprendizaje autónomo puede tomar tiempo, grandes costos computacionales y puede no ser práctico en escenarios de alta complejidad.

El aprendizaje por refuerzo interactivo permite que un entrenador externo de consejos al agente mientras aprende una tarea. En este trabajo, se propone enseñar a un agente de RL a controlar un dron usando las técnicas de reward-shaping y policy-shaping de manera simultánea a través de comandos de voz. Para llevar a cabo el entrenamiento, se plantearon dos escenarios simulados, uno sin obstáculos y otro con obstáculos. Además, se estudió la influencia de cada técnica. Los resultados mostraron que un agente entrenado con ambas técnicas simultáneamente obtiene una recompensa menor que un agente usando solo policy-shaping. Sin embargo, el agente obtiene tiempos de ejecución más bajos y tiene menos dispersión durante el entrenamiento.

Aprendizaje por refuerzo interactivo

En algunas situaciones, permitir que un agente aprenda una tarea completamente de manera autónoma no es práctico debido al alto costo de cada prueba y error. Además, el aprendizaje autónomo envuelve problemas de exploración y una tendencia débil que evita encontrar una política óptima [1]. El aprendizaje por refuerzo interactivo (IRL) es un enfoque que considera un entrenador con conocimiento del ambiente, que da consejo al agente con el fin de optimizar el tiempo de aprendizaje. El consejo puede ser obtenido de un entrenador experto o no, de agentes artificiales con perfecto conocimiento de la tarea: o, de agentes entrenados previamente [2]. En un escenario de IRL, se espera que la interacción entre el entrenador y el agente sea la mínima posible, en caso contrario, el aprendizaje sería supervisado.

En términos de IRL se distinguen dos formas de relación entre el entrenador y el agente. En el primer método, llamado reward-shaping, el entrenador modifica la recompensa que recibe el agente del ambiente. La Fig. 1 muestra el método de reward-shaping, donde el entrenador externo modifica o no la recompensa obtenida del ambiente. La recompensa del entrenador informa como es el desempeño de la acción seleccionada en la iteración anterior [3].

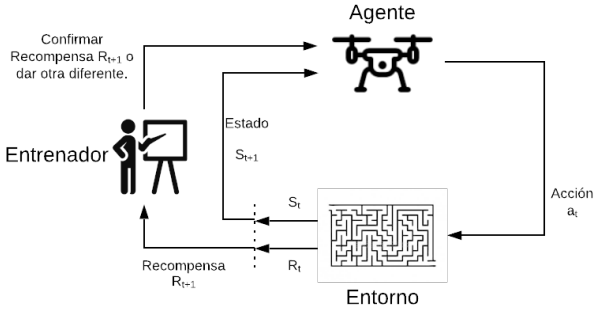

Figura 1

Reward-shaping: Interacción entre el agente y el ambiente, donde un entrenador externo interviene en la recompensa recibida por el ambiente.

El segundo método, llamado policy-shaping, el entrenador modifica la acción seleccionada por el agente y sugiere una acción para ser aplicada. La Fig. 2 muestra el método policy-shaping, donde el entrenador externo propone una nueva acción para ser desarrollada en vez de la acción seleccionada por el agente. En este enfoque se espera que la acción dada por el entrenador tenga un mejor desempeño que la acción del agente, aumentando la probabilidad que sea seleccionada [4].

Cabe resaltar que reward-shaping modifica la estimación del retorno esperado, dando mayor valor a acciones de alto desempeño, pero puede ser problemático para otro tipo de acciones. En policy-shaping la política del agente cambia en vez de afectar la función de recompensa [5], sin embargo, el agente puede demorar en encontrar una política óptima debido a la calidad del consejo [6, 7]. Una comparación entre reward-shaping y policy-sha-

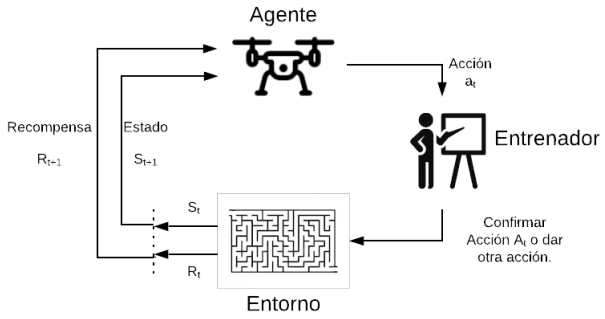

Figura 2

Policy-shaping: Interacción entre el agente y el ambiente, donde un entrenador externo interviene en las acciones seleccionadas por el agente.

ping es realizada por Bignold et al. [8]. En este trabajo los autores muestran que los usuarios que brindan consejos basados en policy-shaping a los agentes brindan consejos más precisos y están dispuestos a ayudar al agente durante más tiempo, brindando más consejos por episodio.

Método

Arquitectura propuesta

En la literatura, el aprendizaje por refuerzo interactivo considera una fuente de variación externa, es decir, un entrenador externo al ambiente que proporciona feedback al agente durante su entrenamiento [9]. En nuestra arquitectura propuesta, el agente recibe feedback de dos fuentes de variación, una bajo reward-shaping y otra bajo policy-shaping. El agente y el ambiente interactúan en cada iteración t. A cada paso, el agente recibe una representación del estado del ambiente, st, seleccionando una acción, at, disponible en el estado actual.

Fig. 3 presenta un diagrama de la arquitectura propuesta, donde el agente interactúa con el ambiente mientras es aconsejado por un entrenador externo bajo los dos enfoques. Una vez seleccionada la acción, el enfoque de policy-shaping se lleva a cabo. Así, con una probabilidad de feedback action La, el entrenador externo selecciona una nueva acción a't, que modifica el estado. Como resultado de llevar a cabo la acción, el agente recibe un valor de recompensa ρt+1 y alcanza un nuevo estado st+1. Una vez que el agente recibe la recompensa, el enfoque de reward-shaping se lleva a cabo. Con una probabilidad de feedback reward Lρ, el agente recibe del entrenador externo una nueva recompensa ρ't.

Por otro lado, se usó el algoritmo de coincidencia de fonemas presentado en [10] para procesar el consejo que da el entrenador externo. El consejo es dado por un audio seleccionado de un conjunto de grabaciones. Para transformar la señal de audio en texto, se utilizó Google Cloud Speech (GCS) en combinación con un lenguaje basado en dominio. Las transmisiones de audio se reciben y se envían al servicio GCS basado en la nube a través de Web Speech API, de donde se obtiene una oración reconocida como hipótesis.

A continuación, se comparó la hipótesis y el diccionario basado en el dominio utilizando la distancia de Levenshtein, seleccionando la instrucción con distancia mínima. Una vez que el comando de voz se convirtió en texto, la señal se procesó y clasificó como una recompensa, en el caso de reward-shaping, y una instrucción para el UAV, en el caso de policy-shaping.

Entorno experimental

Para desarrollar este proyecto, fue usado CoppeliaSim [11], que es un software de simulación de código abierto disponible gratuitamente con una licencia educativa para varios sistemas operativos, como Linux, Windows e iOS, para simular diferentes tipos de robots en entornos realistas. Dos escenarios fueron construidos con el fin de comprobar la metodología propuesta. El primer escenario fue una grilla de 10x10 metros cuadrados, el ambiente no contiene obstáculos permitiendo que el dron tenga trayectorias libres. El segundo escenario es una grilla de 10x10 metros cuadrados, con 11 pilares de 0.80 metros de diámetro distribuidos por todo el ambiente que obstruyen el paso del UAV en algunas trayectorias. La Fig. 4 muestra el escenario con obstáculos en el simulador.

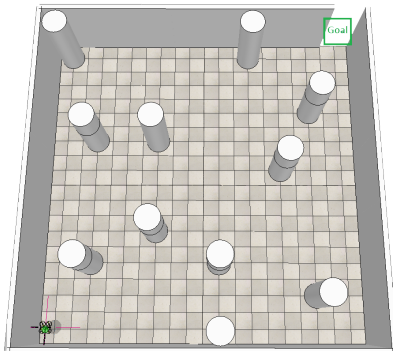

Figura 4

Entorno simulado en CoppeliaSim. En la esquina inferior izquierda se encuentra la UAV, y en la esquina superior derecha se encuentra el objetivo. La tarea consiste en llevar a la UAV por el ambiente hacia el objetivo evitando colisiones con los obstáculos.

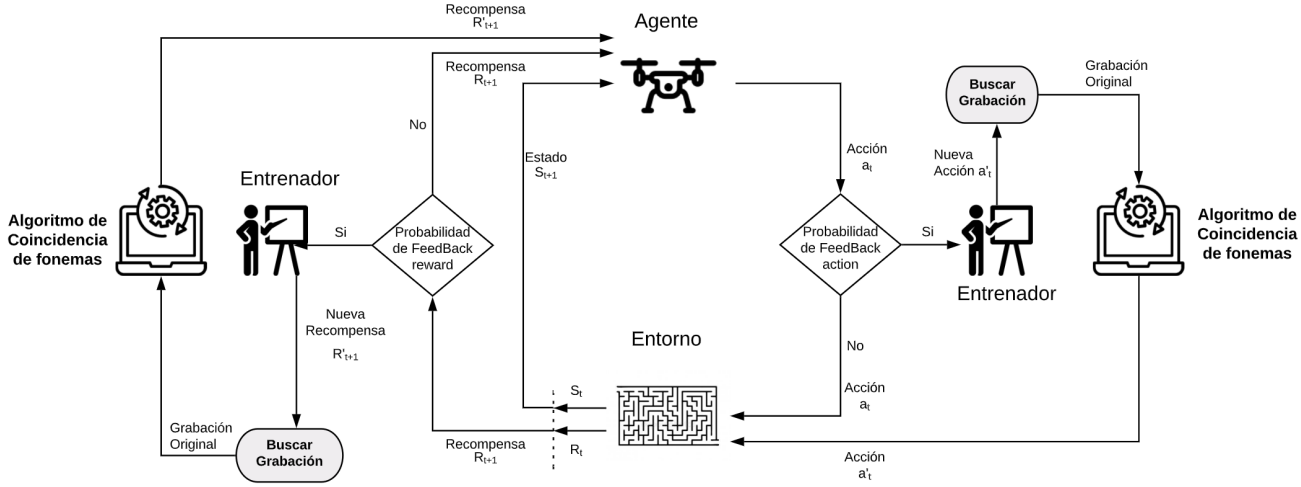

Figura 3

Arquitectura propuesta para el agente de IRL.

Estos escenarios tendrán un UAV al comenzar la simulación (lado inferior izquierdo de la Fig. 4), el cual podrá moverse por toda la superficie, con el fin de salir por la apertura del lado superior derecho (rectángulo verde en la Fig. 4). El dron posee 4 sensores de proximidad con un alcance de 1 metro de distancia, estos se encuentran en los 4 puntos cardinales "Norte", "Sur", "Este" y "Oeste", por lo tanto si existe algún tipo de objeto, ya sea muro o pillar, que se encuentre frente alguno de estos sensores, se puede apreciar un cambio de estado. Con esto, para el escenario sin obstáculos existen 400 estados posibles (10x10x4). Para el escenario con obstáculos existe 356 estados posibles, ya que 11 estados son ocupados por los pilares (10x(10-11)x4).

Para un mejor desempeño de la simulación, fueron impuestos algunos limitantes a los escenarios propuestos:

- La altitud máxima de la unidad aérea no puede ser mayor a 2.50 metros, por lo tanto, si el dron se encuentra a esa altitud, y se le solicita subir, éste no efectuará la acción.

- La altitud mínima de la unidad aérea no puede ser menor a 0.50 metros, por lo tanto, si el dron se encuentra a esa altura, y se le solicita bajar, éste no efectuará la acción.

- Si al dron se le solicita una acción que implique moverse en una dirección, y existe un objeto o muralla que obstruya el paso del UAV, el sensor en esa dirección se activará, lo cual implica que no se efectuará el movimiento del dron.

Con el fin de evaluar el enfoque, fueron propuestos 4 experimentos que se aplicaron en cada uno de los escenarios. Los experimentos tienen las siguientes características:

- 20 agentes RL entrenados autónomamente.
- 20 agentes IRL, los cuales tendrán una probabilidad de consejo de policy-shaping del 15%, donde el entrenador es uno de los agentes entrenados autónomamente en el experimento anterior.
- 20 agentes IRL, los cuales tendrán una probabilidad de consejo de reward-shaping del 15%, donde el entrenador es el mismo agente autónomo que en el experimento anterior.
- 20 agentes IRL, los cuales tendrán una probabilidad de consejo de policy-shaping y reward-shaping del 15%, donde el entrenador es el mismo agente autónomo que en el experimento anterior.
- Cada agente es entrenado durante 20 episodios usando el algoritmo Q-learning

Resultados

En la Fig. 5 se pueden apreciar la recompensa promedio por episodio de los agentes RL, para el escenario sin obstáculos. Se observa que los agentes con una curva de aprendizaje más lenta fueron los agentes autónomos (línea azul). La recompensa promedio de los agentes entrenados con la técnica de policy-shaping (línea naranja) es levemente superior a los agentes autónomos. Sin embargo, la curva permanece constante después del episodio 12 y toma valores similares a los agentes autónomos. Los agentes con las técnicas policy-shaping y reward-shaping simultáneamente (línea roja), obtuvieron una recompensa menor que los agentes mencionados anteriormente, pero superior a los agentes con la técnica reward-shaping, sin embargo, después del episodio 10 permanecen constantes. La mayor recompensa promedio total es alcanzada por los agentes entrenados usando la técnica policy-shaping, esto acorde con lo observado en la Fig. 5.

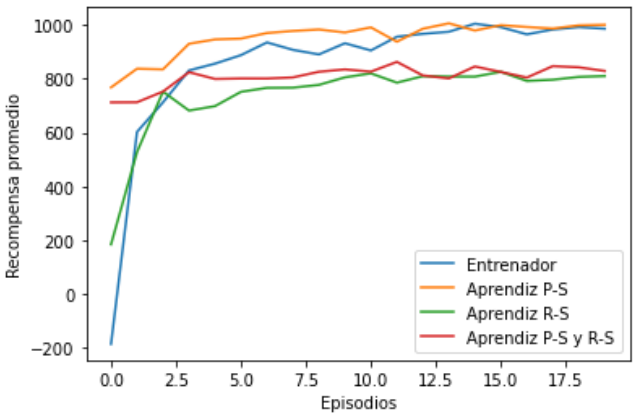

Figura 5

Gráfico de recompensas promedio de los 20 agentes en el escenario sin obstáculos.

Con respecto al escenario con obstáculos (Fig. 6), se observa que los agentes con la técnica policy-shaping obtuvieron una recompensa mayor que los agentes autónomos en los primeros episodios, pero después del episodio 11 alcanzaron valores de recompensa similares. Los agentes entrenados con las técnicas policy-shaping y reward-shaping simultáneamente presentan un comportamiento ligeramente constante, teniendo un comportamiento similar a los agentes entrenados con reward-shaping después de los 5 episodios. Para finalizar, los agentes con policy-shaping tuvieron una recompensa mayor en todos los episodios, sin embargo, los agentes autónomos tienen un comportamiento similar a estos después del episodio 8. La mayor recompensa promedio total es alcanzada nuevamente por los agentes con la técnica policy-shaping, esto acorde con lo observado en la Fig. 6.

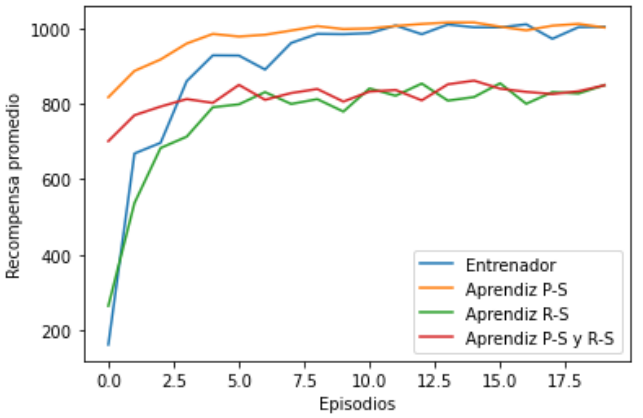

Figura 5

Gráfico de recompensas promedio de los 20 agentes en el escenario con obstáculos.

En general, los agentes entrenados con la técnica reward-shaping son los que tienen menores valores, esto es debido a que su función de recompensa es distinta a los agentes autónomos y los implementados con policy-shaping. Uno de los cambios más significativos es cuando el agente llega al objetivo final, otorgando una recompensa de 800, en comparación con la otorgada a los otros agentes que tiene un valor de 1000. Esta diferencia influye de forma negativa en el comportamiento de las curvas de aprendizaje de los agentes de reward-shaping.

Conclusiones

En este documento se presentó un enfoque de IRL para enseñar agentes de aprendizaje por refuerzo usando los métodos de reward-shaping y policy-shaping. El enfoque propuesto fue implementado en el control de un UAV en un ambiente simulado. Fue evaluada la influencia de los métodos reward-shaping y policy-shaping individual y simultáneamente en el desempeño del agente durante el entrenamiento, también se compararon los tiempos de ejecución de cada experimento.

En términos de recompensa promedio, en el experimento sin obstáculos el agente con policy-shaping obtiene valores altos durante todos los episodios, aunque después del episodio 11 aproximadamente, el agente autónomo obtiene valores similares. Por otro lado, el agente con reward-shaping y policy-shaping simultáneamente obtiene valores de recompensa más altos que el agente autónomo en los primeros episodios, pero no mantiene una tendencia creciente como el autónomo. Respecto al experimento con obstáculos, se observa un comportamiento similar al experimento anterior. En este experimento hay una diferencia notable entre la recompensa promedio de los agentes autónomos y con policy-shaping, con los agentes reward-shaping, y reward-shaping y policy-shaping simultáneamente. Las curvas de recompensa promedio de agentes con policy-shaping tienen una tendencia a subir en los primeros episodios, y aquellas con reward-shaping se mantienen constantes en los episodios finales. Este comportamiento sugiere que un consejo basado en reward-shaping no beneficia al aprendizaje en los primeros episodios, aunque ayuda a mantener un buen desempeño en los episodios finales. Por el contrario, un consejo basado en policy-shaping beneficia el aprendizaje en los primeros episodios.

Para los dos escenarios, los agentes con policy-shaping obtuvieron en promedio una recompensa mayor. Dado que policy-shaping aporta información sobre las acciones y no sobre su desempeño, la dispersión para estos agentes será mejor bajo este método, pero el método reward-shaping puede ayudar a mantener una baja dispersión en los últimos episodios.

Referencias

[1] W Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In Proceedings of the fifth international conference on Knowledge capture. Association for Computing Machinery New York NY United States, Redondo Beach California USA, 9-16.

[2] Adam Bignold, Francisco Cruz, Richard Dazeley, Peter Vamplew, and Cameron Foale. Persistent rule-based interactive reinforcement learning. Neural Computing and Applications (2021): 1-18.

[3] Andrea L. Thomaz and Cynthia Breazeal. 2007. Asymmetric Interpretations of Positive and Negative Human Feedback for a Social Learning Agent. In RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication. IEEE, Jeju, South Korea, 720-725.

[4] Andrea L. Thomaz, Guy Hoffman, and Cynthia Breazeal. 2006. Reinforcement Learning with Human Teachers: Understanding How People Want to Teach Robots. In ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication. IEEE, Hatfield, UK, 352-357.

[5] Adam Bignold, Francisco Cruz, Matthew E Taylor, Tim Brys, Richard Dazeley, Peter Vamplew, and Cameron Foale. 2021. A conceptual framework for externally-influenced agents: An assisted reinforcement learning review. Journal of Ambient Intelligence and Humanized Computing - (2021), 1-24.

[6] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. Advances in neural information processing systems 26 (2013).

[7] Adam Bignold, Francisco Cruz, Richard Dazeley, Peter Vamplew, and Cameron Foale. An evaluation methodology for interactive reinforcement learning with simulated users. Biomimetics 6, no. 1 (2021).

[8] Adam Bignold, Francisco Cruz, Richard Dazeley, Peter Vamplew, and Cameron Foale. 2022. Human engagement providing evaluative and informative advice for interactive reinforcement learning. Neural Computing and Applications (2022), 1-16.

[9] Cristian C Millán-Arias, Bruno JT Fernandes, Francisco Cruz, Richard Dazeley, and Sérgio Fernandes. 2021. A robust approach for continuous interactive actor-critic algorithms. IEEE Access 9 (2021), 104242-104260.

[10] Ruben Contreras, Angel Ayala, and Francisco Cruz. 2020. Unmanned aerial vehicle control through domain-based automatic speech recognition. Computers 9, 3 (2020), 75.

[11] Eric Rohmer, Surya PN Singh, and Marc Freese. 2013. V-REP: A versatile and scalable robot simulation framework. In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems IROS. IEEE, Tokyo, Japan, 1321-1326.