CEA Comité Español de Automática
XLV JORNADAS AUTOMÁTICA MÁLAGA 2024
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Simulación de interacción humano-robot basada en la mirada en entornos virtuales

Menéndez, E.*, Martínez, S., Monje, C. A., Balaguer, C.

RoboticsLab, Dpto. de Ingeniería de Sistemas y Automática, Universidad Carlos III de Madrid, Av. De la Universidad 30, 28911 Leganés, España.

To cite this article: Menendez, E., Martínez, S., Monje, C. A., Balaguer, C. 2024. Simulation of gaze-based human-robot interaction in virtual environments.
Jornadas de Automática, 45. https://doi.org/10.17979/ja-cea.2024.45.10958

Resumen

Este artículo presenta un nuevo método de interacción humano-robot mediante el seguimiento de la mirada en entornos virtuales. Este enfoque reproduce aplicaciones reales en las que el usuario, equipado con gafas de seguimiento ocular, indica al robot qué objetos necesita fijando su mirada. Esta interacción se ha implementado en el simulador Gazebo, donde el usuario controla una cámara flotante con un mando. La cámara flotante imita la vista que ofrecen las gafas de seguimiento ocular y muestra esta perspectiva en la pantalla situada frente al usuario. Además, se instala una cámara dirigida hacia su rostro para determinar la zona de la pantalla que está observando. Utilizando esta información en el método de identificación del umbral de dispersión, se distingue eficazmente entre fijaciones y movimientos sacádicos de la mirada. Los experimentos preliminares realizados demuestran que el sistema es capaz de identificar el objeto en el que el usuario fija su mirada en entornos virtuales.

Palabras clave: Percepción y detección, Interfaces inteligentes, Interacción Humano-Robot, Mirada, Simulación

Simulation of gaze-based human-robot interaction in virtual environments

Abstract

This article presents a new method of human-robot interaction using gaze tracking in virtual environments. This approach replicates real-world applications where a user, equipped with eye-tracking glasses, indicates to the robot which objects they need by fixing their gaze. This interaction has been implemented in the Gazebo simulator, where the user controls a floating camera with a gamepad. This floating camera mimics the view provided by the eye-tracking glasses and displays this perspective on the screen in front of the user. Additionally, a camera aimed at the user's face is installed to determine the area of the screen they are observing. Using this information in the dispersion threshold identification method, the system effectively distinguishes between fixations and saccadic movements of the gaze. Preliminary experiments have demonstrated that the system is capable of identifying the object that the user is gazing at in virtual environments.

Keywords: Perception and sensing, Intelligent interfaces, Human-Robot Interaction, Gaze, Simulation

1. Introducción

La Interacción Humano-Robot (*Human-Robot Interaction*, HRI) es uno de los componentes más importantes de la robótica asistencial, ya que el robot necesita interactuar con las personas y el entorno de manera eficaz. En este contexto, la mirada se destaca como un método de comunicación intuitivo y fácilmente comprensible. Además, la mirada permite una comunicación directa y rápida de las intenciones del usuario, facilitando una interacción más fluida y efectiva. La mirada como medio de interacción es útil en entornos donde el robot debe entregar objetos cuando el usuario los necesite, especialmente cuando se trata de personas con problemas cognitivos o dificultades en el lenguaje. Si el usuario se fija en una zona concreta, el robot debe identificar el objeto deseado

*Autor para correspondencia: emenende@ing.uc3m.es
Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

y entregárselo. Para detectar la zona de atención de la mirada del usuario, algunos trabajos utilizan una cámara externa (Saran et al., 2018), limitándose a escenarios donde tanto el objeto de atención del usuario como su cara deben ser visibles por la cámara. En otros trabajos los usuarios llevan gafas de seguimiento ocular que incluyen cámaras apuntando hacia las pupilas y una cámara adicional para capturar el punto de vista del usuario. La zona de atención del usuario se obtiene proyectando la mirada sobre su punto de vista. Las ventajas de estos métodos son la mayor libertad de movimiento del usuario en el escenario y la precisión en la estimación de la mirada (Valtakari et al., 2021).

El trabajo descrito en (Menendez et al., 2024) desarrolla un nuevo enfoque HRI en el que el robot identifica y recoge objetos basándose exclusivamente en la mirada del usuario, con la ayuda de unas gafas de seguimiento ocular. Este enfoque demostró ser eficiente en diferentes condiciones de visualización sin marcadores externos, posiciones específicas de los objetos o un conocimiento previo sobre los objetos presentes en el escenario. Para abordar este desafío, se utiliza un estimador de posición, orientación y forma basado en supercuádricas utilizando la cámara RGB-D del robot. Además, se utilizaron redes siamesas para relacionar el objeto en el que el usuario fijaba su mirada con el objeto más similar en el campo de visión del robot con el objetivo de identificar el objeto deseado. A continuación, se determinan las posiciones de agarre del objeto deseado, y se planifican y ejecutan los movimientos de alcance para el agarre.

La simulación de la HRI en entornos virtuales proporciona un control detallado sobre el entorno experimental, reduciendo riesgos para usuarios y robots. Además, permite preentrenar las acciones del robot minimizando el tiempo de interacción necesario entre el humano y el robot. En este contexto, (Abal-Fernández et al., 2023) desarrollaron una plataforma de simulación para entrenar sistemas de toma de decisiones en robots asistenciales usando vídeos egocéntricos grabados por usuarios sanos. En este artículo se propone trasladar a un simulador el proceso completo donde el usuario selecciona un objeto con la mirada y el robot lo recoge. El uso de gafas de realidad virtual podría simular la vista de las gafas de seguimiento ocular en entornos virtuales; sin embargo, esta tecnología resulta costosa y necesita software especializado.

Debido a estas limitaciones se ha optado por un sistema más accesible. La Figura 1 muestra las interfaces de interacción con el simulador de las que dispone el usuario. Este puede controlar mediante un mando la posición y orientación de una cámara flotante colocada en el simulador, proporcionando una libertad de movimiento similar a la que ofrecen las gafas de seguimiento ocular en la realidad. Esta configuración permite al usuario explorar el entorno de manera natural, observando una vista en primera persona de la cámara simulada.

La mirada del usuario se detecta mediante una cámara ubicada debajo de la pantalla que el usuario observa. Se utilizan modelos de estimación para mapear el vector de la mirada desde la perspectiva del usuario hacia la pantalla, proyectando así el punto de la mirada en la pantalla. Además, se emplea el método de Identificación por Umbral de Dispersión (*Dispersion-Threshold Identification*, I-DT) para diferenciar los movimientos sacádicos de las fijaciones en zonas específicas del entorno simulado.

El articulo se organiza de la siguiente manera: En la Sección 2 se discuten las diferentes etapas de la estrategia utilizada en la HRI real. La Sección 3 describe el sistema de HRI basado en la mirada en el entorno de simulación. La Sección 4 muestra los experimentos realizados y los resultados obtenidos. Finalmente en la Sección 5, se presentan las conclusiones y las lineas futuras de este trabajo.

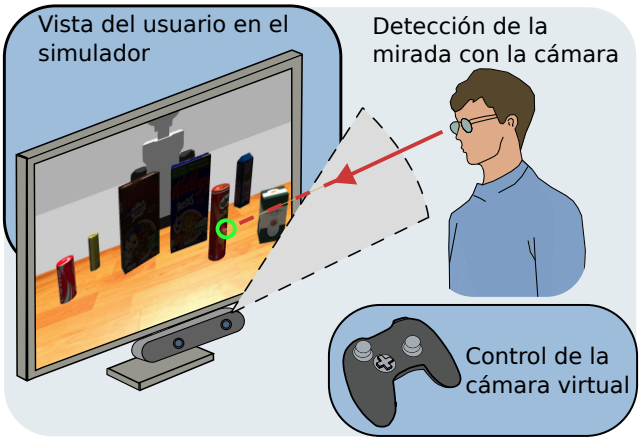


Figura 1: Sistema de interacción con el entorno simulado.

2. Identificación y agarre de objetos basado en la mirada del usuario

En esta estrategia, descrita en mayor detalle en (Menendez et al., 2024), el robot TIAGo++ (Pages et al., 2016) se encuentra observando una mesa con varios objetos, y el objetivo del usuario es seleccionar uno de ellos con la mirada. El procedimiento consiste en un sistema multi-etapa para determinar el objeto en el que el usuario ha fijado su mirada dentro del sistema de referencia del robot, seguido del agarre del objeto por parte del robot (Figura 2).

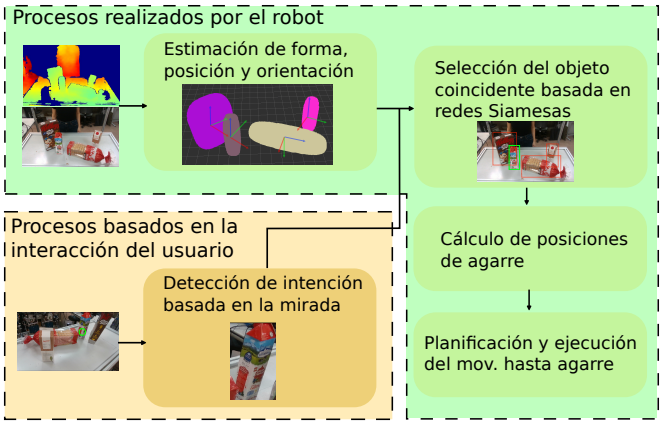


Figura 2: Diagrama del proceso en el sistema HRI real.

El primer proceso consiste en estimar la forma, la posición y la orientación de los objetos utilizando el sensor RGB-D del robot. Primero, se transforma la imagen de profundidad en una nube de puntos 3D, eliminando la superficie horizontal para aislar los objetos. Es importante destacar que esta nube de puntos puede presentar áreas vacías por oclusiones, cuando partes del objeto no son visibles desde el ángulo de la cámara

o están ocultas por otros objetos. Después, la nube de puntos se segmenta en *clusters*, que pertenecen a objetos distintos. Por último, se estiman las supercuádricas para cada uno de los *clusters*. Estas son representaciones geométricas modeladas matemáticamente a partir de un conjunto de parámetros que describen de manera simplificada la forma, posición y orientación de los objetos.

Simultáneamente, el usuario comienza mirando a una zona alejada de los objetos para prevenir una fijación precipitada. El usuario puede navegar el entorno libremente y fijarse en cualquier objeto cuando desee. El módulo de intención basado en la mirada opera en tiempo real analizando puntos de fijación de la mirada obtenidos mediante las gafas de seguimiento ocular. Dicho proceso proporciona la intención de agarrar un objeto, junto con una probabilidad de decisión. Cuando la probabilidad supera un umbral, el proceso proporciona un recorte del objeto desde el punto de vista del usuario.

Cuando se recibe el recorte del objeto en el que el usuario ha fijado su mirada, el sistema lo compara con la vista desde la perspectiva del robot para identificar el objeto deseado. En este proceso, se emplea una red siamesa entrenada mediante tripletas de imágenes y la función *triplet loss* para reconocer similitudes entre diferentes perspectivas del mismo objeto. El entrenamiento se realiza con tripletas de imágenes: una imagen ancla del objeto desde el punto de vista del usuario, una imagen positiva del mismo objeto desde la perspectiva del robot y una imagen negativa de un objeto diferente visto por el robot. La función *triplet loss* trabaja para minimizar la distancia euclidiana entre los vectores de características de las imágenes ancla y positiva, mientras maximiza la distancia entre la ancla y la negativa. Durante la aplicación práctica, se obtienen las *bounding boxes* de los objetos en la imagen del robot utilizando las supercuádricas, y la red siamesa compara el recorte del objeto desde la perspectiva del usuario con los recortes de los objetos en la imagen del robot, identificando el objeto con la menor distancia euclidiana entre los vectores de características.

El siguiente paso es determinar las posiciones de agarre, identificando el eje más largo de la supercuádrica como el principal. El gripper se posiciona perpendicularmente a este eje y su apertura se ajusta a la dimensión del objeto en esa dirección. Se consideran múltiples posiciones de agarre a lo largo del eje principal, así como posibles rotaciones del gripper. Las posiciones de agarre se establecen con respecto a un sistema de coordenadas situado en el centro de la supercuádrica y se transforman al sistema del robot.

El último proceso consiste en planificar y ejecutar el movimiento de agarre. A partir de las posiciones de agarre, se generan configuraciones articulares deseadas para ambos brazos del robot. Utilizando una versión modificada del algoritmo RRT bidireccional (Menéndez et al., 2024), (Haustein et al., 2019), se calcula el movimiento de cada brazo, seleccionando la trayectoria más óptima para agarrar el objeto deseado. El algoritmo construye un árbol RRT *forward* desde la posición de reposo del brazo y múltiples árboles *backward* desde las configuraciones articulares, seleccionando la trayectoria de menor coste al comparar constantemente las soluciones encontradas.

3. Simulación de la Interacción Humano-Robot basada en la mirada

En este artículo, se traslada todo la estrategia al simulador Gazebo (Koenig and Howard, 2004), ampliamente utilizado en robótica. Los procesos realizados por el robot se trasladan directamente al simulador (Figura 2), sin necesidad de modificaciones, aprovechando la disponibilidad del robot simulado en Gazebo. Sin embargo, la simulación del sistema de interacción basado en la mirada no es directa. Adaptar los procesos de HRI basado en la mirada a un entorno simulado plantea desafíos únicos, especialmente en asegurar la libertad de movimiento y la precisión en la detección de la mirada dentro de un mundo virtual. Este sistema de interacción simulado incluye elementos críticos como el control de la cámara del usuario en el simulador a través de un mando, el cálculo de las *bounding boxes* de los objetos desde la perspectiva de dicha cámara, y la identificación precisa de las fijaciones de la mirada del usuario en un objeto específico.

3.1. Control de la Cámara del Usuario

Se ha creado una cámara flotante en Gazebo con una resolución de 1280x720 y 30 fps, especificaciones que garantizan una calidad de imagen suficientemente alta para simular la percepción visual humana y asegurar una experiencia fluida en el simulador. El control de la posición de la cámara se realiza mediante el mando utilizando el paquete de ROS (Robot Operating System) joy (Koubâa et al., 2017), que lee los datos de los botones y estados de los *joysticks*. El software desarrollado transforma estas lecturas en movimientos de cámara deseados en el simulador.

La Figura 3 ilustra cómo se utiliza el mando para controlar la cámara flotante en el simulador. La Figura 4(a) muestra la cámara flotante en el simulador. El *joystick* derecho permite mover la cámara hacia adelante/atrás y derecha/izquierda en los ejes locales x e y, mientras que el *joystick* izquierdo permite rotar la cámara alrededor de los ejes locales z e y. La cruceta se utiliza para mover la cámara hacia arriba/abajo en el simulador. Además, se requiere mantener pulsado el botón de hombre muerto para activar los controles, una medida de seguridad común en robótica. La Figura 4(b) muestra el punto de vista de la cámara flotante, que simula la vista en primera persona del usuario en el simulador. Esta vista permite al usuario explorar el entorno virtual de manera natural y emular la libertad de movimiento que proporcionan las gafas de seguimiento ocular.

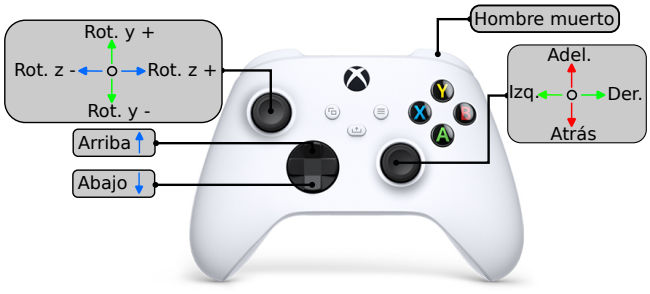
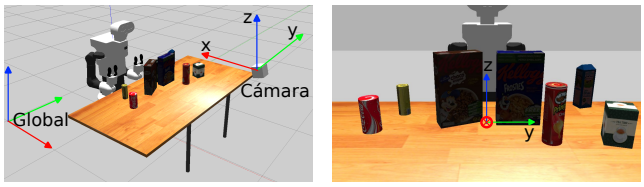


Figura 3: Controles del mando para manejar la cámara flotante.



(a) (b)

Figura 4: (a) Escenario del simulador con la cámara flotante. (b) Vista en primera persona del usuario en el simulador.

3.2. *Obtención de bounding boxes en la vista de la cámara del usuario*

Para obtener las *bounding boxes* 2D de los objetos en la vista de la cámara flotante del simulador, se ha optado por evitar el uso de detectores de objetos basados en redes neuronales, que requieren un entrenamiento extenso y están limitados a un conjunto específico de objetos predefinidos (Redmon et al., 2016). En su lugar, se ha optado por un proceso que emplea la información que proporciona el simulador Gazebo sobre todos los modelos dispuestos en el escenario.

Se ha desarrollado un plugin para Gazebo que proporciona a través de un servicio de ROS la forma básica de los modelos (cajas, cilindros o esferas), sus dimensiones, y sus posiciones y orientaciones en el sistema de referencia global. Utilizando esta información, se obtienen las *bounding boxes* 3D orientadas de cada modelo, representadas por sus vértices. Los vértices de las bounding boxes 3D en el sistema global (${}^{Global}\mathbf{p}_i$) se transforman al sistema de la cámara controlada por el usuario. Conocidas la posición y orientación de dicha cámara en tiempo real, los vértices de las bounding boxes 3D en el sistema de la cámara (${}^{Cámara}\mathbf{p}_i$) se pueden obtener con la Ecuación 1.

$$ {}^{Cámara}\mathbf{p}_i = ({}^{Global}\mathbf{H}_{Cámara})^{-1} \cdot {}^{Global}\mathbf{p}_i \tag{1}$$

Siendo ${}^{Global}\mathbf{H}_{Cámara}$ la matriz de transformación desde el sistema de referencia global al sistema de la cámara.

La Ecuación 2 se emplea para transformar los vértices a las coordenadas de la imagen usando el modelo de la cámara de *pinhole*:

$$\begin{pmatrix} u_i \\ v_i \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} {}^{C}x_i \\ {}^{C}y_i \\ {}^{C}z_i \end{pmatrix} \tag{2}$$

donde f_x y f_y son las distancias focales en los ejes x e y, y (x_0, y_0) es el punto principal. Las bounding boxes 2D en la imagen se calculan encontrando el mínimo y máximo de las coordenadas transformadas en ambos ejes. Específicamente, la bounding box se representa como $(u_{min}, v_{min}, u_{max}, v_{max})$ en píxeles, donde la esquina superior-izquierda se corresponde con los valores mínimos (u_{min}, v_{min}), y la esquina inferior-derecha se corresponde con los valores máximos (u_{max}, v_{max}). Estas bounding boxes se utilizan para detectar en qué objeto se está fijando el usuario.

3.3. *Detección de la fijación de la mirada del usuario*

Una vez que la vista del usuario en el entorno simulado está configurada y controlada, el siguiente paso es detectar a qué punto de la pantalla está mirando el usuario. Esto se realiza mediante una cámara Intel RealSense configurada a 720x480 y 60 fps, situada debajo del monitor que se encuentra frente al usuario. La tasa de fps afecta a la frecuencia a la que se estiman y proyectan los puntos de mirada en la pantalla en el punto de vista del usuario, asegurando una estimación en tiempo real.

Aunque la cámara utilizada contiene un sensor de profundidad, se ha optado por un método que emplea las imágenes RGB para estimar la mirada del usuario. Este método, descrito en (Falch and Lohan, 2024), utiliza modelos de estimación de la mirada basados en la apariencia. La metodología se basa en la detección de la cara del usuario mediante una cámara RBG y el uso de redes neuronales convolucionales (Convolutional Neural Network, CNN) para estimar el vector de la mirada. El modelo proporciona un vector unitario de la mirada que se proyecta en la pantalla del ordenador, permitiendo determinar el punto exacto donde el usuario está mirando. Esta técnica elimina la necesidad de equipos adicionales o marcadores, utilizando solo una cámara RGB y procesos de calibración simples para lograr una estimación precisa del punto de fijación.

El sistema de estimación de la mirada incluye un proceso de calibración antes de mostrar en la pantalla por primera vez el punto de vista del usuario en el simulador. Durante la calibración, se presentan varios puntos en la pantalla que el usuario debe mirar fijamente. La cámara captura imágenes de las pupilas mientras este sigue los puntos de calibración. Utilizando estos datos, el sistema calcula los vectores de la mirada y los mapea en las coordenadas de la pantalla, ajustando los modelos de estimación para mejorar la precisión. Realizando esta calibración previa se asegura que se puede determinar con mayor exactitud el punto de la pantalla al que el usuario está mirando. Una vez que se determina en tiempo real este punto, el siguiente paso es distinguir correctamente los movimientos oculares para que el sistema responda de forma adecuada a la intención del usuario. Los movimientos oculares más comunes son: los movimientos sacádicos, que son desplazamientos rápidos entre fijaciones en el campo visual, y los movimientos involuntarios de escasa amplitud que se producen durante el mantenimiento de la fijación (Pannasch et al., 2008).

Para identificar cuándo el usuario está fijando su mirada, se emplea el método de identificación por umbral de dispersión (Dispersion-Threshold Identification, I-DT) (Salvucci and Goldberg, 2000). Este método evalúa la dispersión de los puntos de la mirada dentro de una ventana temporal específica. Se compara la distancia máxima entre todos los puntos de mirada dentro de esta ventana con un umbral predefinido. Si la dispersión de los puntos de la mirada es menor que el umbral de dispersión, se considera que el usuario está fijando su mirada en esa región de la pantalla. Adicionalmente, para asegurar la precisión, se requiere que durante todos los puntos de la ventana temporal utilizada para obtener la fijación, la cámara controlada por el usuario no esté en movimiento. Además, se calcula la media entre todos los puntos de la ventana empleando dicha media como punto de fijación. Cuando se detecta que el usuario se fija en una zona de la imagen se verifica si este punto de fijación se encuentra dentro de alguna de las *bounding boxes* de los objetos (Figura 5).

Figura 5: Puntos de mirada del usuario (rojos) durante una ventana I-DT y su media (verde). El bounding box verde indica el objeto seleccionado, que contiene la media de los puntos de mirada.

Si este es el caso, se procederá a hacer el proceso completo en el cual el robot mediante redes siamesas comprobará cual de los objetos que tiene delante es más similar al objeto en el que se ha fijado el usuario, si el nivel de parecido sobrepasa un umbral se asume que ese es el objeto que desea el usuario y el robot procederá a recoger el objeto.

4. Experimentos y resultados

En este artículo se han realizado experimentos para definir los parámetros del método I-DT utilizado para la distinción entre los movimientos sacádicos y los movimientos involuntarios que se producen durante la fijación de la mirada en una zona, así como para evaluar la precisión del sistema de detección de fijación de la mirada en el entorno simulado. En la Figura 6 se muestra a un usuario realizando los experimentos, utilizando un mando para mover la cámara en el simulador, con una cámara que estima su mirada y un punto de fijación mostrado en la pantalla.

Figura 6: Un usuario controla la cámara del simulador mediante un mando, mientras una cámara captura y estima su mirada, mostrando el punto de fijación estimado en la pantalla.

4.1. *Definición de los parámetros del método I-DT*

El objetivo del primer experimento es determinar los parámetros óptimos para el método de I-DT. El usuario se coloca frente a una pantalla en la que se muestran cinco imágenes diferentes, cada una de las cuales mostrando múltiples objetos en ubicaciones distintas sobre una superficie horizontal. Además, se instala una cámara mirando hacia la cara del usuario para detectar los puntos de la pantalla a los que está mirando. En este escenario se pide al usuario que se fije en objetos distintos. Para mejorar la precisión del sistema de estimación de la mirada se realiza una fase inicial de calibración con el usuario. Durante esta fase, se le pide al usuario mirar fijamente cuatro puntos situados cerca de las esquinas de la pantalla. Luego, se analizaron los datos obtenidos en estos escenarios para configurar los parámetros del método I-DT: la ventana temporal y la dispersión máxima. Se probaron diversas ventanas temporales para determinar el intervalo más corto en el que los puntos de mirada del usuario mostraban una agrupación consistente alrededor de cada objeto. Al mismo tiempo, se evaluó la dispersión máxima que aún permitía identificar una fijación precisa. El análisis de los resultados reveló que los parámetros óptimos para la ventana temporal y la dispersión máxima fueron de 300 ms y 6° de visión, respectivamente.

4.2. *Evaluación de la detección de fijación de la mirada en objetos*

En este experimento se muestran cinco escenarios distintos en Gazebo, cada uno conteniendo cinco objetos distribuidos aleatoriamente. Los objetos se colocan sin estar apilados sobre una superficie plana. Se instala una cámara apuntando hacia el usuario para estimar su mirada en la pantalla, en la cual se muestra la vista del usuario en primera persona en el simulador. El usuario emplea un mando para mover la cámara flotante del simulador, controlando su vista con total libertad. A medida que el usuario va moviendo la cámara, las *bounding boxes* se van adaptando a dicha vista. Además, para prevenir selecciones involuntarias al comienzo de la prueba, se requiere que el usuario mueva la cámara antes de que el sistema empiece a detectar fijaciones. El método I-DT con los parámetros definidos en el experimento anterior se utiliza para distinguir los movimientos sacádicos de pequeños movimientos involuntarios que se realizan mediante la fijación. Cuando el sistema detecta una fijación dentro de una *bounding box*, esta se resalta en verde, y se solicita al usuario confirmar si el objeto resaltado es efectivamente el objeto de su interés mediante dos botones distintos del mando (Figura 7(a)). Sin embargo, si el sistema detecta una fijación cerca de una *bounding box*, esta se resaltaba en rojo y se le pregunta al usuario si su intención era mirar ese objeto, a lo que el usuario responde afirmativa o negativamente (Figura 7(b)).

Durante la realización de las pruebas en los cinco escenarios, se recolectó información de un total de 120 fijaciones. A partir de estas interacciones, se calcularon dos métricas: la precisión de identificación de objetos y la tasa de falsos positivos en fijaciones cercanas. La precisión de identificación de objetos mide la proporción de veces que las fijaciones del usuario caen exactamente dentro de las *bounding boxes* de los objetos que intentaba observar y que el usuario ha confirmado como su intención. Dicha métrica alcanzó un 97.5 %, reflejando la capacidad del sistema para identificar correctamente el objeto que el usuario está mirando. Por otro lado, la tasa de falsos positivos en fijaciones cercanas mide la proporción de fijaciones que, pese a estar cerca de una *bounding box* pero no dentro de ella, son indicadas por el usuario como intentos de seleccionar el objeto correspondiente. En un 8.2 % de los casos, el usuario intenta seleccionar un objeto, pero sus fijaciones no caen exactamente dentro de la *bounding box*. Esta tasa elevada puede deberse a que los usuarios a menudo miran hacia los bordes de los objetos, especialmente en el caso

de objetos pequeños, donde pequeñas desviaciones en la mirada pueden resultar en fijaciones fuera de la *bounding box*. Además, la precisión de la calibración del sistema de seguimiento de la mirada afecta a la diferencia entre la ubicación real del punto de fijación y el estimado.

(a)

(b)

Figura 7: Detección de fijaciones: (a) La fijación dentro de una *bounding box* se resalta en verde y se solicita confirmación. (b) La fijación cerca de una *bounding box* se resalta en rojo y se pregunta por la intención.

5. Conclusiones

Este artículo presenta un sistema de interacción humano-robot basado en la detección de la mirada en un simulador. El sistema permite al usuario mover una cámara virtual mediante un mando, cuya perspectiva en primera persona se muestra directamente en la pantalla situada frente a él. Para estimar la zona de la pantalla que el usuario está mirando, se emplea una cámara orientada hacia su rostro, facilitando así una interacción precisa y natural. Los resultados experimentales confirman la capacidad del sistema para interpretar correctamente la fijación del usuario en los objetos situados en el entorno simulado. De cara a futuros trabajos, se plantea la extensión del estudio a múltiples usuarios, replicando todo el proceso en simulación, desde la identificación del objeto fijado por la mirada hasta su recogida por el robot. Estos experimentos tienen como objetivo validar la eficacia del sistema en un entorno virtual más complejo y variado.

Agradecimientos

La investigación que ha conducido a estos resultados ha recibido financiación del proyecto COMPANION-CM: Inteligencia artificial y modelos cognitivos para la interacción simétrica humano-robot en el ámbito de la robótica asistencial, con referencia Y2020/NMT-6660, financiado por Proyectos Sinérgicos de I+D la Comunidad de Madrid.

Referencias

Abal-Fernández, S., Caramazana-Zarzosa, C., Loureiro-Casalderrey, M. B., Martínez, S., Balaguer, C., Díaz-de María, F., González-Díaz, I., 2023. Learning RL policies for anticipative assistive robots by simulating human-robot interactions in real scenarios using egocentric videos. In: 2023 IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 1–8.
DOI: 10.1109/ROBIO58561.2023.10354837

Falch, L., Lohan, K. S., 2024. Webcam-based gaze estimation for computer screen interaction. Frontiers in Robotics and AI 11, 1369566.
DOI: 10.3389/frobt.2024.1369566

Haustein, J. A., Hang, K., Stork, J., Kragic, D., 2019. Object placement planning and optimization for robot manipulators. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 7417–7424.
DOI: 10.1109/IROS40897.2019.8967732

Koenig, N., Howard, A., 2004. Design and use paradigms for gazebo, an open-source multi-robot simulator. In: 2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566). Vol. 3. Ieee, pp. 2149–2154.
DOI: 10.1109/IROS.2004.1389727

Koubâa, A., et al., 2017. Robot Operating System (ROS). Vol. 1. Springer, ISBN: 978–3–319-54926-2.

Menéndez, E., Martínez, S., Balaguer, C., 2024. Selección y agarre robótico de objetos basada en el seguimiento de la mirada. In: Actas del Simposio de Robótica, Bioingeniería y Visión por Computador. Universidad de Extremadura, Servicio de Publicaciones, pp. 127–131,
ISBN: 978-84-9127- 262-5.

Menendez, E., Martínez, S., Díaz-de María, F., Balaguer, C., 2024. Integrating egocentric and robotic vision for object identification using siamese networks and superquadric estimations in partial occlusion scenarios. Biomimetics 9 (2).
DOI: 10.3390/biomimetics9020100

Pages, J., Marchionni, L., Ferro, F., 2016. Tiago: the modular robot that adapts to different research needs. In: International workshop on robot modularity, IROS. Vol. 290.

Pannasch, S., Helmert, J. R., Roth, K., Herbold, A.-K., Walter, H., 2008. Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. Journal of Eye Movement Research 2 (2).
DOI: 10.16910/jemr.2.2.4

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788.
DOI: 10.48550/arXiv.1506.02640

Salvucci, D. D., Goldberg, J. H., 2000. Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the 2000 symposium on Eye tracking research & applications. pp. 71–78.
DOI: 10.1145/355017.355028

Saran, A., Majumdar, S., Thomaz, A., Niekum, S., 2018. Real-time human gaze following for human-robot interaction. In: Proceedings of the International Conference on Human Robot Interaction.
DOI: 10.1109/IROS.2018.8593580

Valtakari, N. V., Hooge, I. T., Viktorsson, C., Nyström, P., Falck-Ytter, T., Hessels, R. S., 2021. Eye tracking in human interaction: Possibilities and limitations. Behavior Research Methods, 1–17.
DOI: 10.3758/s13428-020-01517-x