Simposio CEA de Robótica, Bioingeniería, Visión Artificial y Automática Marina 2025

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

Sistema de detección de usuarios y comunicación proactiva para un robot autónomo de servicio

Antonio Cañete*, Cipriano Galindo, Francisco-Ángel Moreno, Javier González-Jiménez

Grupo de Percepción Artificial y Robótica Inteligente (MAPIR), Dept. de Ingeniería de Sistemas y Automática.

Instituto Universitario en Ingeniería Mecatrónica y Sistemas Ciberfísicos (IMECH.UMA).

Universidad de Málaga, Blvr. Louis Pasteur, 35, 29071 Málaga, España.

Resumen

La navegación segura es un aspecto crítico para los robots sociales que se desplazan en entornos concurridos. Además, estos robots deben respetar los espacios personales para evitar generar desconfianza y, al mismo tiempo, interactuar de forma efectiva y socialmente aceptable. Este trabajo presenta un sistema multimodal que combina técnicas de visión por computador, planificación de trayectorias socialmente aceptables y estrategias de comunicación proactiva con las personas cercanas al robot. El sistema ha sido validado en escenarios controlados, mostrando resultados preliminares prometedores en términos de robustez y fluidez social.

Palabras clave: Robótica autónoma, Interacción humano-robot, Robótica inteligente, Sistemas de percepción y sensores, Sistemas multiagente, Sistemas de navegación

User detection system and proactive communication for an autonomous service robot.

Abstract

Safe navigation is a critical aspect for social robots moving in crowded environments. In addition, these robots must respect personal space to avoid creating distrust and, at the same time, interact in an effective and socially acceptable way. This paper presents a multimodal system that combines advanced computer vision techniques, socially-aware trajectory planning and proactive communication strategies with the people near the robot. The system has been validated in controlled scenarios, showing promising preliminary results in terms of robustness and social smoothness.

Keywords: Autonomous robotics, Human-robot interaction, Intelligent robotics, Perception and sensing systems, Multi-agent systems, Navigation systems

1. Introducción

En las últimas décadas, la *robótica de servicio*, robots concebidos para ejecutar tareas útiles para las personas fuera de la cadena industrial, ha pasado de ser un nicho experimental a un mercado consolidado a escala global con una rápida penetración en áreas como logística, salud y asistencia social (World Robotics 2024 – Service Robots, 2025). Este auge ha dado lugar a la aparición de los llamados *robots sociales*, diseñados para interactuar de manera natural con usuarios humanos en múltiples aplicaciones, tales como asistencia en entornos hospitalarios, acompañamiento a personas mayores (Blindheim et al., 2022) o guía de visitantes en museos y ferias (Rosa et al., 2024).

En este contexto, la capacidad de *detectar y reconocer* usuarios, así como de ofrecer *comunicación proactiva* por parte del robot, constituye un pilar esencial para alcanzar interacciones fluidas y eficaces. La mera navegación autónoma, basada en algoritmos de localización y mapeo (SLAM) y en planificadores de rutas, resuelve la problemática de la navegación segura, aunque no aborda plenamente la necesidad de comprender y anticipar las intenciones de los usuarios. De ahí que sea fundamental *combinar técnicas de visión por computador* (detección de pose, reconocimiento de rostros, seguimiento de personas), *proce-*

*Autor para correspondencia: antbaena@uma.es

samiento de audio (análisis de señales de voz, localización de emisores) y *planificación de comportamientos* que permitan al robot tomar la iniciativa cuando detecte que un usuario requiere asistencia o desea interactuar.

El objetivo principal de este trabajo es el desarrollo de un sistema de detección de usuarios y comunicación proactiva para un robot de servicio que opere en entornos compartidos con usuarios.

El éxito de la interacción natural del robot depende de la latencia y el tiempo de ejecución de todos los componentes que lo componen, por lo que en este artículo se valora el despliegue del mismo en arquitecturas de computación en el borde (*edge computing*) (Ambrosio-Cestero et al., 2024).

En las siguientes secciones, se describe el diseño de la arquitectura propuesta y la implementación de los módulos de detección de personas, procesamiento de sonido y navegación social. Se detallan asimismo las estrategias de integración en la plataforma robótica elegida y los resultados preliminares de validación en un entorno real.

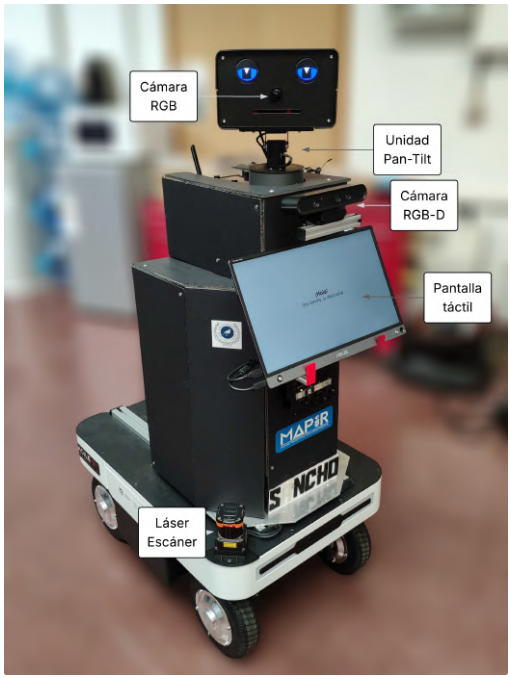

Figura 1: Robot de servicio Sancho. Está equipado con dos escáneres láser hokuyo usados para localización y navegación, y sensores RGB y RGB-D para detección de usuarios.

2. Descripción general del sistema

El sistema propuesto se ha desarrollado usando el framework ROS2 Humble y se ha implantado en el robot móvil Sancho (ver figura 1).

2.1. Plataforma robótica y sistema de sensorización

El robot Sancho está formado por una base robótica AgileX Ranger Mini V3, diseñada para realizar operaciones en entornos interiores y semiexteriores. Sobre la base se han incorporado dos escáneres láser Hokuyo, ubicados en esquinas diagonales del chasis: uno en la esquina frontal derecha y otro en la trasera izquierda, ambos girados 45º con respecto al eje principal del robot.

Se incorpora una cámara RGB-D Orbbec Astra, montada en la parte superior del robot. Dicha cámara permite estimar la posición y distancia de los usuarios en el entorno, así como ajustar la trayectoria en entornos dinámicos. Además, se incluye una cámara RGB instalada en la cabeza robótica, enfocada a la interacción social.

Esta configuración permite una navegación eficiente y suave, otorgando al robot capacidades reactivas y seguras frente a cambios repentinos en el entorno. El uso combinado de los sensores (láser, RGB-D y RGB) favorece la identificación temprana de usuarios u obstáculos y la posterior adaptación de la trayectoria, reforzando la fiabilidad de la operación en escenarios concurridos.

2.2. Arquitectura software y orquestación de módulos

La arquitectura modular del sistema permite la interacción dinámica entre navegación e interacción social. Para la detección de usuarios, se emplea la red MoveNet (TensorFlow, 2024) y procesamiento de audio basado en técnicas de localización de la fuente sonora (Rocha et al., 2021). El sistema identifica la presencia y posición de personas, respeta normas de proxemia mediante una capa de costo personalizada y planifica rutas seguras a su alrededor.

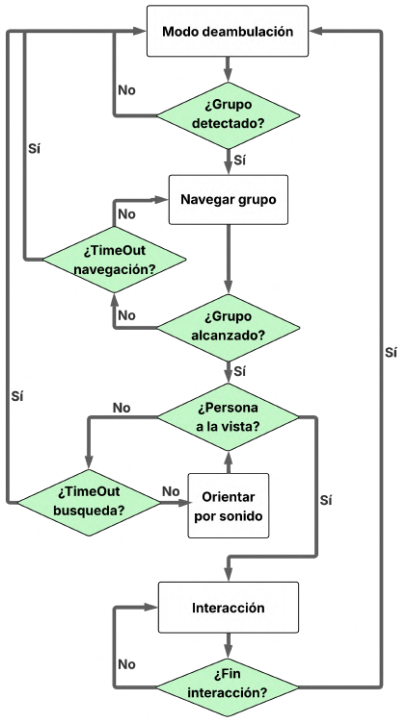

Figura 2: Comportamiento social. El robot deambula por su entorno hasta detectar a usuarios o grupos, con los que intenta interactuar.

La figura 2 ilustra el comportamiento implementado. Cuando el robot detecta a un grupo de personas, establece un nuevo punto de navegación en donde puede iniciar un proceso de interacción. Se utiliza la cámara RGB para la detección facial y orientación de la cabeza hacia las personas, y se emite un mensaje predefinido al usuario, reforzando la sensación de comunicación personalizada. Un orquestador central supervisa de

forma continua las condiciones del entorno y gestiona dinámicamente la transición entre los estados de búsqueda, aproximación e interacción, activando o desactivando los módulos correspondientes en cada fase. Este mecanismo asegura un comportamiento social coherente y fluido, al tiempo que permite definir rutinas personalizadas para cada tarea.

3. Detección de usuarios

Es clave que, para una navegación social correcta, el robot sea capaz de detectar a las personas y estimar su posición en el entorno. De esta manera, no solo puede ubicar puntos de navegación adecuados para transmitir mensajes proactivamente, sino también ajustarse a criterios de proxemia y desplazarse de forma respetuosa entre los usuarios.

3.1. Detección de usuarios con MoveNet

A partir de las imágenes RGB, el sistema emplea MoveNet MultiPose, cuyo modelo cuantizado ronda los 3-20 MB y se mantiene estable por encima de 25 fps en CPU de gama media. Los keypoints obtenidos por MoveNet se asocian a la imagen de profundidad, lo que permite estimar la pose 3D del torso de cada usuario, información crítica para la planificación de trayectorias socialmente aceptables.

Otras alternativas que se han barajado han sido OpenPose (Cao et al., 2017) y MediaPipe (Lugaresi et al., 2019). OpenPose requiere una GPU de ≥ 4 GB VRAM para superar los 0,3 fps que ofrece utilizando únicamente CPU, quedándose por debajo de 15 fps. Por su parte, MediaPipe, si bien es muy ligero, sigue sin ofrecer una detección multiusuario robusta en escenarios de concurrencia.

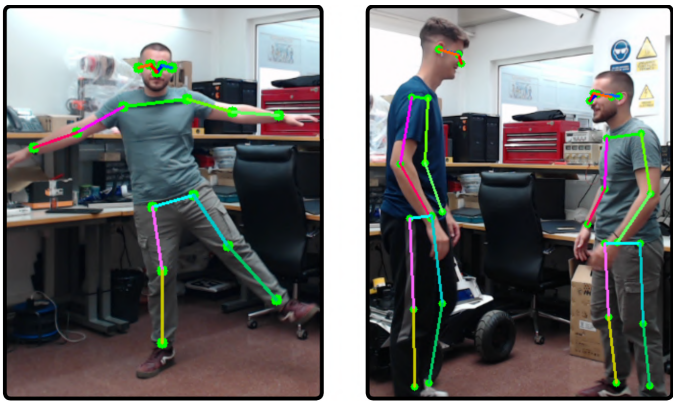

Figura 3: Detección de usuarios. La red Movenet detecta a los usuarios y permite calcular su orientación.

3.2. Agrupación y detección de grupos

En entornos concurridos, el sistema usa DBSCAN (Ester et al., 1996) para agrupar automáticamente a las personas detectadas según la proximidad de sus keypoints. DBSCAN define dos parámetros esenciales: el radio de vecindad ϵ y el número mínimo de vecinos *minPts*. Para equilibrar sensibilidad y robustez se fijan por defecto *minPts* = 2 puntos y ϵ = 1,0 m, lo que evita que detecciones esporádicas formen clústeres ficticios; ambos valores pueden ajustarse *on-the-fly* mediante los parámetros dinámicos de ROS 2 para adaptarse a distintas densidades de público. Así, el robot puede decidir si debe acercarse a una persona o a un grupo, ajustando la trayectoria para mantener la distancia en cada caso.

3.3. Integración para la navegación social

La información sobre las personas y los grupos detectados se publica en ROS2 como objetivos que el módulo de navegación consume para actualizar el mapa de obstáculos y planificar rutas socialmente aceptables. Cada individuo o grupo se modela mediante una distribución gaussiana asimétrica. Esta representación permite al planificador generar rutas que evitan aproximaciones bruscas o invasivas.

Asimismo, la aparición de nuevos grupos activa el módulo de interacción proactiva, que ordena al robot desplazarse a un punto próximo al conjunto identificado para transmitir el mensaje correspondiente. De este modo, la detección de usuarios y la comunicación proactiva confluyen, proporcionando una interacción más fluida y natural en entornos compartidos.

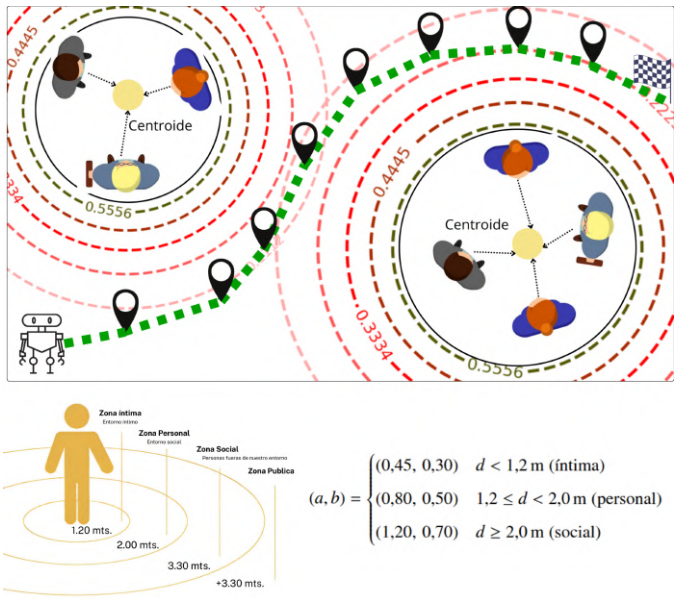

Figura 4: Navegación socialmente aceptada. Tanto para grupos como para individuos se establecen diferentes distancias sociales representadas por elipses. Los semiejes (a, b) de dichas elipses se establecen en los valores indicados en la figura para cada caso.

4. Navegación socialmente aceptada

Para un robot que comparte espacio con humanos, no es suficiente que solo evite obstáculos, es necesario que su movimiento respete las normas socioculturales que regulan la proxemia (Hall, 1966). Con ese fin, el sistema extiende el stack de navegación con una capa de coste social que modela y penaliza la invasión de los espacios interpersonales.

4.1. Modelado proxémico mediante elipses anisotrópicas

Cada usuario detectado en el campo de visión del robot genera una zona de influencia que se representa mediante una elipse cuyo centro coincide con la proyección en el plano del torso y cuyo eje mayor se alinea con la orientación corporal θ.

Aunque, en teoría, podríamos ajustar dinámicamente los semiejes de la elipse (a, b) en función de múltiples variables como la edad del interlocutor, el contexto cultural o densidad de ocupación, en este sistema se emplean unos valores fijos que garantizan un comportamiento socialmente aceptable en entornos interiores. Se consideran los valores para respetar el espacio personal, es decir, $a = 0{,}8\,\text{m}$ (dirección de la mirada) y $b = 0{,}5\,\text{m}$ (lateral y posterior). Estos valores hacen que el robot mantenga unos márgenes suficientes para iniciar la interacción sin invadir el espacio personal, a la vez que reduce la penalización lateral para favorecer adelantamientos y maniobras en espacios estrechos. La sustitución de estos parámetros por otros pertenecientes a la lista de distancias (íntima o social) puede realizarse en tiempo de ejecución si se requiere un ajuste más preciso en situaciones específicas.

El interior de la elipse se modela como un campo escalar gaussiano asimétrico $C(x, y)$:

$$C(x,y) = \exp\!\Big(-\tfrac{1}{2}\,\mathbf{p}^{\mathsf{T}} R_\theta\, \Sigma^{-1} R_\theta^{\mathsf{T}} \mathbf{p}\Big),$$

$$\mathbf{p} = \begin{pmatrix} x - x_c \\ y - y_c \end{pmatrix}, \qquad \Sigma = \operatorname{diag}(a^2,\, b^2).$$

donde R_θ es la matriz de rotación que gira Σ un ángulo θ para alinear la distribución con la orientación del usuario. De este modo, las trayectorias de aproximación, ya sean laterales, frontales o posteriores, reciben penalizaciones diferentes, lo que favorece un comportamiento percibido como no invasivo.

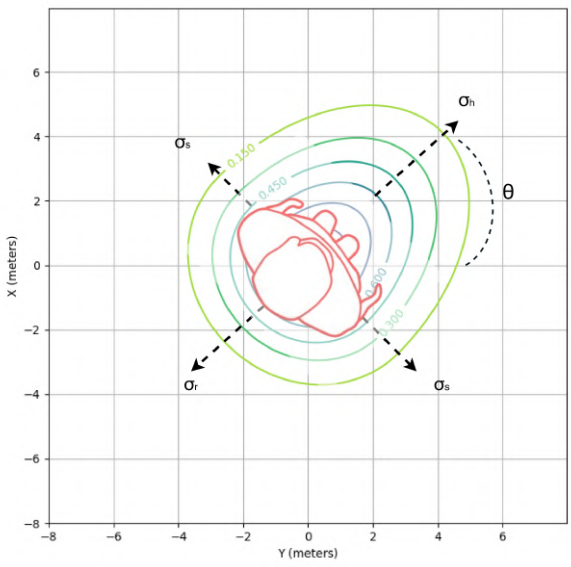

Figura 5: Campo escalar gaussiano asimétrico alrededor de una persona, alargado en la dirección frontal y rotado un ángulo θ respecto al eje Y, correspondiente a la orientación corporal.

5. Sistema de comunicación proactiva

Tras recibir la señal de fin de navegación emitida por el orquestador cuando el robot alcanza la posición de interacción cercana al grupo previamente detectado, y el sistema conmuta al modo de interacción. En este estado se detiene la navegación y se ejecuta el siguiente flujo de control:

1. **Activación sensorial**. Se habilitan la cámara RGB frontal y los dos micrófonos estéreo incorporados dentro de la cámara RGB-D astra.
2. **Verificación y búsqueda multimodal**.
   a) *Intento visual*. Se ejecuta una detección facial rápida sobre los primeros fotogramas. Si se detecta al menos un rostro que se estime relativamente cercano al robot por su tamaño en la imagen, se pasa directamente al paso 3.
   b) *Estimación acústica guiada*. Si no se encuentra ningún rostro, el robot comienza a utilizar los micrófonos para calcular la dirección dominante del sonido aplicando Correlación Generalizada con Ponderación de Fase (GCC-PHAT) (Knapp and Carter, 1976), ya que este método, con solo un par de micrófonos, ofrece estimaciones robustas ante ruido. El método consiste en normalizar la fase del espectro en ambos canales, calcular la diferencia de tiempo de llegada (TDOA) y, a partir de ella, determina la dirección dominante del sonido $\hat{\theta}_{\text{mic}}$
   c) *Barrido local*. Utilizando el ángulo de la fuente de sonido, $\hat{\theta}_{\text{mic}}$, el robot comienza a girar la cabeza (o el chasis, si el ángulo fuera demasiado grande) explorando alrededor de la posición prevista del grupo. Durante todo el proceso se repite la detección facial, si aparece un rostro, se continúa el flujo; de lo contrario, se sigue el bucle.
   d) *Cancelación*. Si transcurre t_{search} sin hallar usuario, el sistema da por perdida la interacción, replanificando la misión y volviendo al estado inicial.
3. **Detección facial**. Cada fotograma I_t es procesado por el detector basado en HOG-SVM proporcionado por `DLIB` (King, 2009). El resultado se expresa como un conjunto de rostros acotados de la forma $b_i = (x, y, w, h, s)$, donde s es la confianza en la detección.
4. **Selección del interlocutor principal**. Entre las detecciones válidas se elige la persona cuya posición facial esté más cerca del centro de la imagen, minimizando la distancia euclidiana entre su rostro y el centro de la escena:

$$i^\star = \arg\min_i \left(\sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \right),$$

   donde (x_i, y_i) representa la coordenada del centro facial de la persona i, y (x_c, y_c) corresponde al centro geométrico de la imagen.
5. **Orientación (gaze shifting)**. A partir de las coordenadas de $b_{i^\star}$ se calcula el vector de error en píxeles $(\Delta u, \Delta v)$ y se convierte a ángulos de pan–tilt mediante

$$\alpha_x = \arctan\!\left(\frac{\Delta u}{f_x}\right), \quad \alpha_y = \arctan\!\left(\frac{\Delta v}{f_y}\right),$$

   con f_x, f_y extraídos de la matriz de calibración intrínseca $\mathbf{K}$. Estos ángulos se utilizan para dirigir la unidad pan–tilt al centro del rostro.

6. **Emisión y seguimiento simultáneos**. Mientras se reproduce el audio, el detector permanece activo; tan pronto como se detecta un rostro, este se marca como posible interlocutor principal al que la unidad pan–tilt podría orientarse utilizando (α_x, α_y) sin interrumpir la locución.
7. **Criterio de finalización**. El estado de interacción termina cuando concluye el mensaje y se agota un segundo temporizador T_{timeout}

6. Arquitectura de implementación

El sistema se ha implementado íntegramente sobre el *framework* de ROS2 Humble. Cada capacidad del robot —percepción visual, audio, navegación y comunicación— se implementa como nodos independientes con una clara interfaz de entrada-salida. Esta descomposición favorece la sustitución de algoritmos o modelos y el escalado de los módulos que más carga generan.

6.1. *Stack de navegación: Nav2*

La navegación autónoma se realiza gracias al framework de navegación **Navigation2** (Nav2) (Macenski et al., 2023, 2020). Nav2 ofrece servidores independientes para planificación global, control local, suavizado, gestión del mapa y orquestación mediante *Behavior Trees*, lo que facilita comportamientos sociales avanzados y la inserción de una capa de coste proxémica personalizada, descrita en la Sección 3.3.

6.2. *Despliegue distribuido sobre la plataforma CSAR*

Para el despliegue se emplea CSAR (Cloud System Architecture for Robotics) (Ambrosio-Cestero et al., 2024), la cual proporciona contenedores Linux persistentes, de baja latencia y con acceso directo a GPU. Cada nodo con alta demanda computacional se despliega de forma independiente sobre CSAR, lo que permite procesar operaciones complejas en tiempo reducido. Gracias a su arquitectura, CSAR facilita la migración eficiente de nodos con alto coste computacional en el robot hacia entornos optimizados con acceso acelerado a hardware, reduciendo significativamente la latencia en la ejecución de tareas críticas.

7. Resultados preliminares

Las pruebas iniciales se realizaron en el *laboratorio de MAPIR* de la E.T.S. de Ingeniería Informática (superficie $\approx 100\,\text{m}^2$), representando un entorno de interiores normalmente concurrido. El robot *Sancho* patrulló rutas aleatorias con obstáculos estáticos (mesas, sillas) y personas en movimiento simulando trayectorias cruzadas. Cada sesión constó de un recorrido inicial de "deambulación" para luego pasar a navegación "social" con parada e interacción verbal.

El robot logró *bordear* a usuarios individuales manteniendo la distancia interpersonal mínima ($> 0{,}75\,\text{m}$) y ejecutó acercamientos a cada grupo, respetando las elipses proxémicas generadas por la capa de costo. La Figura 6 muestra dos fotogramas anotados:

Los resultados confirman la robustez perceptual del sistema en interiores bien iluminados, con tasas de detección superiores al 90 %. Las condiciones de baja iluminación penalizan significativamente la precisión visual dada la menor calidad de las imágenes RGB, pero no comprometen la seguridad de navegación gracias al respaldo de los escáneres láser.

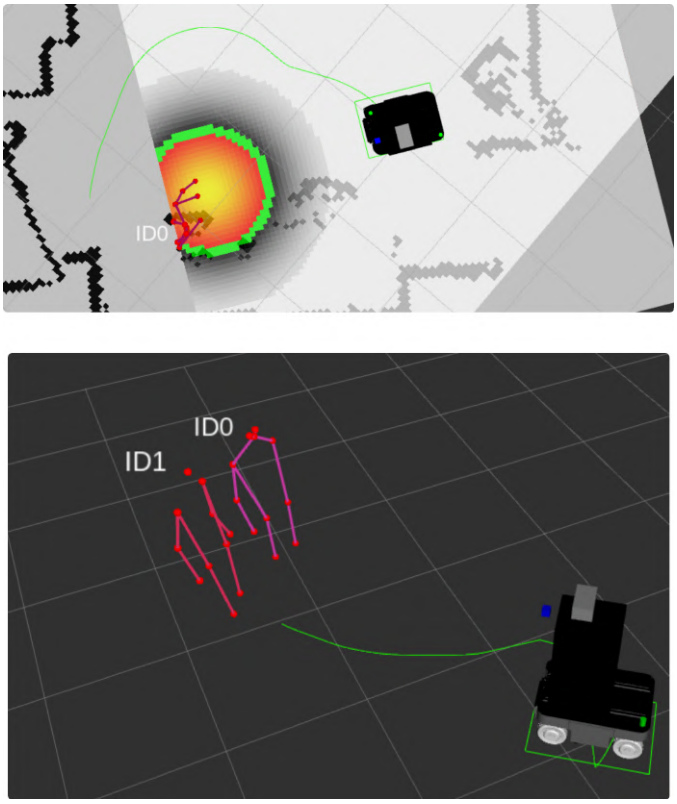

Figura 6: Arriba) Navegación sin intención comunicativa. El robot bordea al usuario respetando su distancia personal. Abajo) Acercamiento socialmente aceptado a un grupo de personas con las que se va a interactuar.

8. Conclusiones

Este trabajo ha descrito el diseño e implementación de un sistema de navegación social capaz de detectar usuarios e iniciar tareas de comunicación proactiva para robots de servicio. La propuesta integra:

- Una planificación de trayectorias enriquecida con capas de coste proxémicas, capaces de mantener distancias interpersonales acordes con la teoría de la proxemia;
- Un orquestador de contextos que alterna de forma autónoma entre los modos de navegación e interacción;
- Módulos de percepción multimodal: detección de pose humana mediante MoveNet y localización visual y acústica del interlocutor, que permiten al robot identificar al usuario dominante y dirigir hacia él su mirada y sus mensajes de voz;

La validación funcional, realizada en escenarios reales, confirmó que el robot preserva la consistencia social de su comportamiento: respeta el espacio personal de los usuarios durante el desplazamiento, evita aproximaciones bruscas y orienta su cámara al interlocutor correcto, reforzando la percepción de interacción personalizada.

Agradecimientos

Este trabajo ha sido desarrollado en el contexto de los proyectos MINDMAPS (PID2023-148191NB-I00) y Voxeland (JA.B1-09), financiados por el Ministerio de Ciencia e Innovación y la Universidad de Málaga, respectivamente.

Referencias

Ambrosio-Cestero, G., Matez-Bandera, J. L., Ruiz-Sarmiento, J. R., González-Jiménez, J., 2024. Entorno basado en contenedores linux para el desarrollo de aplicaciones robóticas. In: Jornadas de Automática. Vol. 45.

Blindheim, K., Solberg, M., Hameed, I. A., Alnes, R. E., 2022. Promoting activity in long-term care facilities with the social robot Pepper: A pilot study. Informatics for Health and Social Care.
DOI: 10.1080/17538157.2022.2086465

Cao, Z., Simon, T., Wei, S.-E., Sheikh, Y., July 2017. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7291–7299.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd. Vol. 96. pp. 226–231.

Hall, E. T., 1966. The Hidden Dimension. Doubleday, New York.

King, D. E., 2009. Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research 10, 1755–1758.

Knapp, C., Carter, G., 1976. The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech, and Signal Processing 24 (4), 320–327.
DOI: 10.1109/TASSP.1976.1162830

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., Grundmann, M., 2019. Mediapipe: A framework for building perception pipelines. CoRR abs/1906.08172.
URL: http://arxiv.org/abs/1906.08172

Macenski, S., Martín, F., White, R., Ginés Clavero, J., 2020. The marathon 2: A navigation system. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).
URL: https://github.com/ros-planning/navigation2

Macenski, S., Moore, T., Lu, D. V., Merzlyakov, A., Ferguson, M., 2023. From the desks of ros maintainers: A survey of modern & capable mobile robotics algorithms in the robot operating system 2. Robotics and Autonomous Systems.

Rocha, G. D., Torres, J. C. B., Petraglia, M. R., Vorländer, M., 2021. Direction of arrival estimation of partial sound sources of vehicles with a two-microphone array. Acta Acustica 5, 18.
DOI: https://doi.org/10.1051/aacus/2021011

Rosa, S., Randazzo, M., Landini, E., Bernagozzi, S., Sacco, G., Piccinino, M., Natale, L., 2024. Tour guide robot: a 5g-enabled robot museum guide. Frontiers in Robotics and AI 10, 1323675.

TensorFlow, 2024. MoveNet: Ultra fast and accurate pose detection model. Último acceso: 2025-05-02.
URL: https://www.tensorflow.org/hub/tutorials/movenet

World Robotics 2024 – Service Robots, 2025. Executive Summary, accessed April 2025.
URL: https://ifr.org/wr-service-robots/