CEA Comité Español de Automática

XLV JORNADAS AUTOMÁTICA MÁLAGA 2024

UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Sistema multimodal para la orientación de robots móviles hacia su interlocutor

Cañete, A.*, Quemada-Torres, E.*, Ruiz-Sarmiento, J.R., Moreno, F.A., González-Jiménez, J.

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.

To cite this article: Cañete, A., Quemada-Torres, E., Ruiz-Sarmiento, J.R., Moreno, F.A., González-Jiménez, J. 2024. Multimodal system for the orientation of mobile robots towards its interlocutor. Jornadas de Automática, 45. https://doi.org/10.17979/ja-cea.2024.45.10939

Resumen

Con el objetivo de lograr una interacción humano-robot lo más natural posible, es fundamental que el robot se oriente hacia su interlocutor. Este trabajo presenta un sistema multimodal que usa información visual y de sonido para lograr una orientación precisa incluso en situaciones complejas con múltiples personas, personas fuera del campo de visión del sensor, etc. En concreto, un sistema de micrófonos estéreo es el encargado de detectar el inicio y fin de la interacción, así como de calcular el ángulo de incidencia del sonido para iniciar la orientación del robot. Por su parte, la información visual proveniente de una cámara se usa para localizar la presencia del interlocutor mediante detección facial, asistida por el ángulo de incidencia del sonido. Una vez localizado, el sistema se encarga de orientarse constantemente hacia dicha persona de manera precisa. El trabajo incluye una demostración del comportamiento del sistema en escenarios límite utilizando el robot social Sancho.

Palabras clave: Robótica inteligente, Aprendizaje automático, Interacción multimodal, Integración y percepción de sensores, Sistemas de control del movimiento

Multimodal system for mobile robot orientation towards its interlocutor

Abstract

To facilitate natural human-robot interaction, it is crucial for the robot to orient itself towards its interlocutor. This work presents a multimodal system that uses visual and sound information to achieve precise orientation even in complex scenarios with multiple people, people outside the sensor's field of view, etc. Specifically, a stereo microphone system detects the start and end of the interaction, as well as calculates the angle of incidence of the sound to initiate the robot's orientation. Visual information from a camera is used to locate the presence of the interlocutor through face detecion, assisted by the angle of incidence of the sound. Once located, the system precisely and constantly orients itself towards that person. The work includes a demonstration of the system's performance in edge cases using the social robot Sancho.

Keywords: Intelligent robotics, Machine Learning, Multi-modal interaction, Sensor integration and perception, Motion Control Systems

1. Introducción

La Interacción Humano-Robot (del inglés *Human-Robot Interaction*, *HRI*) es un campo de estudio que busca desarrollar sistemas robóticos capaces de interactuar con las personas de una manera natural y efectiva (Sheridan, 2016). Un aspecto clave de esta interacción es la orientación del robot hacia la persona con la que está interactuando, lo que afecta significativamente a la calidad de la comunicación y la percepción de naturalidad en el comportamiento del robot.

Un enfoque primario a la hora de orientar al robot hacia su interlocutor consiste en utilizar imágenes provenientes de cámaras montadas a bordo para, en una primera etapa, detec-

*Autor para correspondencia: antbaena@uma.es, eulogioquemada@uma.es

tar la cara de la persona, y en una segunda, orientar al robot en base a la posición de dicha detección en la imagen. Si la cámara se encuentra montada en el eje central vertical del robot, el objetivo consiste en realizar una rotación que centre la posición de la cara en la imagen. En el estado del arte pueden encontrarse multitud de técnicas de detección facial que podrían emplearse para este fin, desde las más tradicionales pero con un alto rendimiento basadas en el clasificador de Viola-Jones (Ruiz-Sarmiento et al., 2011), hasta las más sofisticadas que emplean técnicas de aprendizaje profundo (Baltanas-Molero et al., 2021). Aunque bastante efectivo, este enfoque presenta limitaciones. Por ejemplo, sería incapaz de orientar al robot en situaciones donde el interlocutor se encuentre fuera del campo de visión de la cámara, o en escenarios donde haya más de una persona dentro de dicho campo, lo que requeriría procedimientos adicionales para identificar la que se encuentra interactuando con el robot.

Un enfoque alternativo consiste en utilizar una configuración de dispositivos de sonido que permita estimar el ángulo de incidencia del sonido α, como pueda ser un sistema de micrófonos estéreo (Rocha et al., 2021). Una vez recuperado este ángulo, para orientar al robot sólo necesitamos rotarlo en la dirección adecuada hasta convertir α en 90º, lo que indica que la persona se encuentra de frente. De nuevo, aunque efectivo, este enfoque resulta poco preciso, se ve altamente influenciado por el ruido y el sonido ambiente y, en el caso de querer orientar un robot que cuente con más de un grado de libertad con respecto a la parte a posicionar (p. ej., una cabeza robótica montada sobre una unidad *pan-tilt*), el audio estéreo determina el ángulo de incidencia con respecto a un solo eje.

Este trabajo presenta un sistema multimodal que usa tanto información visual como sonido estéreo para orientar al robot hacia su interlocutor. Este enfoque combinado permite superar las deficiencias individuales de cada modalidad. En resumen, el par estéreo es el encargado de detectar cuándo la persona ha iniciado una interacción, calculando en dicho caso el ángulo de incidencia del sonido entrante, e iniciando una rotación del robot para orientarse hacia ella. Para solventar la falta de precisión y la posible indeterminación en algunos ejes, la información visual es procesada en busca de detecciones faciales en ángulos que sean coherentes con el anteriormente estimado. Una vez detectada la cara del interlocutor, el posicionamiento del robot se refina hasta llevar esta al centro de la imagen. De este modo, la combinación de ambas fuentes de información ofrece una solución robusta para la orientación del robot, aumentando el número de escenarios donde es aplicable y mejorando su precisión, permitiendo una respuesta más natural y efectiva. Para ilustrar el funcionamiento del sistema, implementado en ROS2 (Macenski et al., 2022), se hace uso del robot social Sancho, y se describen varios casos de uso en escenarios límite. En concreto se controla la orientación de su cabeza, la cual se encuentra montada sobre una unidad *pan-tilt* que permite posicionarla.

2. Descripción general del sistema

La Fig. 1 ilustra el flujo de trabajo o *pipeline* del sistema. Esta descripción asume que el robot está equipado con una cámara y un sistema de sonido estéreo consistente en dos micrófonos orientados en la misma dirección paralela al suelo y situados a la misma altura. Además, la cámara y el par de micrófonos se han de encontrar centrados horizontalmente, siendo necesario un procesamiento adicional de los ángulos estimados de no ser así. En el instante inicial, el sistema se encuentra en un estado de **no-interacción**, a la espera de que el módulo de detección de interacción indique que una posible interacción se ha iniciado (ver Sec. 3). Esto ocurre cuando se detecta un sonido que cumple con un cierto patrón. Este módulo también se encarga de realizar una estimación aproximada del ángulo de incidencia de la fuente del sonido, esto es, del posible interlocutor. Está estimación se emplea para iniciar el movimiento de orientación hacia la persona.

Una vez comenzada la orientación, el módulo de detección facial procesará las imágenes provenientes de la cámara en busca de detecciones cuya posición horizontal sea coherente con el ángulo de incidencia del sonido (ver Sec. 4). En el caso de alcanzar la orientación de destino sin detectar ningún rostro (p.ej., en situaciones donde la persona no se encuentre visible, haya ocurrido una falsa detección, etc.) el sistema se mantendrá en el estado de no-interacción y quedará a la espera de nuevas activaciones del módulo de detección de interacción. Por contra, si se detecta una cara en una posición coherente, el sistema pasa a un estado interno de **interacción**. El módulo de detección facial cuenta con una fase de reconocimiento o identificación de la persona. Esto permite etiquetar la cara detectada con su identificador único, el cual se usa para orientarse hacia esta persona de manera inequívoca en escenarios con múltiples detecciones faciales. Una vez el interlocutor ha sido localizado, su posición en la imagen se usa para realizar una orientación fina y constante del robot hacia este (ver Sec. 5), ya que el proceso de detección se mantiene en el tiempo.

Podemos destacar que, en este estado, el robot empleará el módulo de detección de interacción para verificar que esta no ha sido concluida. Para ello se comprueba que los ángulos de incidencia del sonido estimados por este módulo siguen

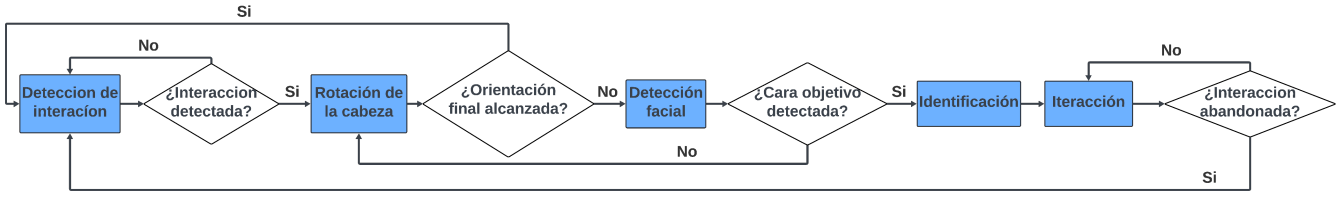

Figura 1: Diagrama de flujo del comportamiento general del sistema.

siendo coherentes con la posición del interlocutor, o si la interacción se ha abandonado. Si esto ocurriese, el sistema volvería al estado de no-interacción.

3. Detección de interacción mediante sonido

El módulo de detección de interacción es el encargado de identificar posibles interacciones por parte del usuario y estimar su orientación aproximada con respecto al robot. Esto se realiza en tres procesos principales (ver Fig. 2): la detección de voz humana (ver Sec. 3.1), la medición de la diferencia de tiempo en la que dicha voz es percibida por cada uno de los micrófonos (ver Sec. 3.2), y la utilización de dicha diferencia para estimar aproximadamente su ángulo de incidencia (ver Sec. 3.3).

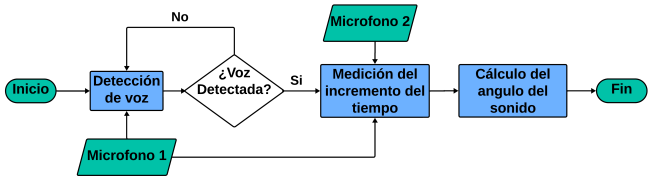

Figura 2: Diagrama de flujo del módulo de detección de interacciones.

3.1. Detección de voz

El proceso de detección de voz trata de discernir cuándo un sonido entrante se trata efectivamente de voz humana. Para ello se hace uso del sonido captado por un único micrófono. Dicho proceso se inicia al percibirse un sonido con una intensidad superior a un cierto umbral. Tras este disparador, el sistema comienza a concatenar las tramas de audio entrantes hasta que bien la intensidad sea inferior al umbral anterior o bien transcurra un cierto tiempo límite. Tras esto, el fragmento de audio resultante es procesado por el modelo de detección de voz humana o VAD (de sus siglas en inglés *Voice Activity Detection*) *pyannote* (Bredin and Laurent, 2021; Bredin et al., 2020). Dicho modelo resuelve si efectivamente se trata de una voz humana, con lo cual se procedería con el siguiente proceso del *pipeline*, o si se descarta el fragmento de audio y se vuelve al paso de detección en caso contrario.

3.2. Medición del incremento del tiempo

Para calcular la variación del tiempo de llegada de la onda de sonido a los diferentes micrófonos, información necesaria para estimar el ángulo de incidencia de dicho sonido, se realiza un proceso de Correlación Cruzada Normalizada (CCN) sobre las ondas percibidas por ambos micrófonos. La CCN emplea el audio representando dichas ondas, desplazando uno sobre otro mientras se aplica la correlación normalizada. Supóngase que t_r^i es una trama del audio capturado por el micrófono derecho, y que t_l^j es una trama capturada por el micrófono izquierdo, entonces esta operación resulta:

$$\text{Correlación}(t_r^i, t_l^j) = \sum_{k=0}^{N-1} t_r^i[k] \times t_l^j[k+m] \tag{1}$$

$$\text{CCN}(t_r^i, t_l^j) = \frac{\text{Correlación}(\mathbf{t_r^i}, \mathbf{t_l^j})}{\|\mathbf{t_r^i}\| \cdot \|\mathbf{t_l^j}\|} \tag{2}$$

De este modo, por ejemplo, la trama t_r^i se desplaza sobre las tramas t_l^j, $j \in [1, |t_l|]$ mientras se aplica la CCN. Aquella correlación con un valor máximo indicará que las ondas capturadas en dichas tramas son las más similares, permitiendo calcular el incremento de tiempo dado su desplazamiento $\Delta(i, j)$ y una frecuencia de muestreo f conocida:

$$\Delta t = \frac{\Delta(i, j)}{f} \tag{3}$$

3.3. Estimación del ángulo de incidencia

Una vez calculado el desplazamiento temporal, es posible calcular el ángulo de incidencia del sonido α empleando el desplazamiento Δt junto con la velocidad del sonido v y la distancia entre los micrófonos b. La Fig. 3 ilustra los distintos elementos involucrados en este cálculo. El ángulo se obtiene de la forma:

$$d_l^2 = (b/2 + d\cos\alpha)^2 + (d\sin\alpha)^2$$

$$d_r^2 = (-b/2 + d\cos\alpha)^2 + (d\sin\alpha)^2$$

$$d_l^2 - d_r^2 = 2bd\cos\alpha \tag{4}$$

Suponiendo $d_l + d_r \approx 2d$, esta ecuación resulta:

$$\cos\alpha = \frac{\Delta d}{b} = \frac{\Delta t \times v}{b} \tag{5}$$

$$\alpha = \arccos\frac{\Delta t \times v}{b} \tag{6}$$

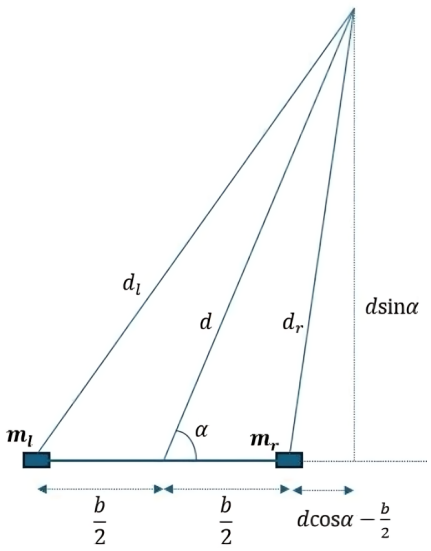

Figura 3: Elementos involucrados en la detección del ángulo de incidencia del sonido empleando un par de micrófonos estéreo.

4. Detección e identificación facial en imágenes

El módulo de detección facial se encarga de localizar la cara del interlocutor en imágenes provenientes de la cámara, así como de asignarle un identificador que permita orientar al robot hacía el interlocutor de manera inequívoca a lo largo del tiempo (ver Fig. 4). Para conseguir esto se llevan a cabo los procesos de detección facial (ver Sec. 4.1), normalización de las caras detectadas para que sigan unos estándares comunes (Sec. 4.2), codificación de dichas caras en forma de un vector de características (Sec. 4.3), y clasificación de dicho vector como perteneciente a una cierta persona (identificación) (Sec. 4.4).

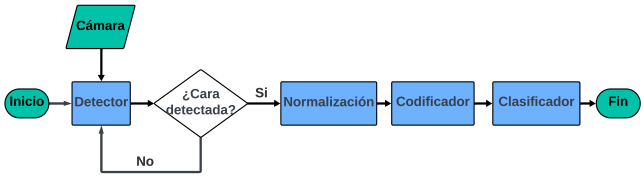

Figura 4: Diagrama de flujo del módulo de detección y reconocimiento facial.

4.1. Detección

La detección facial es un problema ampliamente estudiado, por lo que existen multitud de algoritmos y módulos que realizan esta tarea (Baltanas-Molero et al., 2020). Para el sistema propuesto se ha optado por el detector de caras proporcionado por la librería DLIB (King, 2009), ya que obtuvo un gran rendimiento al someterse a pruebas con el dataset MAPIR Faces (Baltanas-Molero et al., 2021). Otra técnica que consiguió buenos resultados fue el detector MTCNN (Zhang et al., 2016), pero este requería una cantidad de recursos computacionales superior. En cualquier caso, el sistema no está cerrado al uso de un detector concreto, siendo empleable cualquier técnica que use imágenes y devuelva un vector de caras detectadas.

4.2. Normalización

Para aumentar el éxito en la etapa de clasificación, es necesario realizar un proceso de normalización que estandarice en cierta medida el aspecto (orientación, tamaño, brillo y contraste) de las regiones detectadas como caras. Para ello también se recurre a la librería DLIB, en concreto a su sofisticado algoritmo para alineamiento de caras. Este algoritmo se basa en la detección de puntos clave de la cara para calcular la orientación con la que aparece en la imagen, permitiendo aplicar una rotación de forma que la línea que une ambos ojos quede paralela al eje horizontal. Junto a este alineamiento, también se escala cada imagen a un cierto tamaño fijo, correspondiente a un punto intermedio entre el tamaño máximo y mínimo de las caras que son detectables. Por último, empleando el espacio de color YCrCb, se realiza una normalización del histograma de la capa Y, lo que asegura que las distintas caras tendrán unos valores de brillo y contraste similares.

4.3. Codificador

Una vez las detecciones han sido normalizadas, se procede a codificarlas en forma de vector de características. Existen multitud de características, tanto de alto como de bajo nivel, que pueden emplearse para describir caras. Si se considera, por ejemplo, una cara en un recorte de dimensiones 50x50 píxeles, dicha cara podría representarse como un punto en un espacio de 2500 dimensiones. Hipotéticamente, las caras pertenecientes a la misma persona deberían aparecer cercanas en dicho espacio, y se podría emplear un clasificador para decidir si una nueva cara pertenece a una u otra persona. No obstante, este espacio trivial resulta en un gran número de dimensiones con un bajo poder discriminante, que es a su vez altamente sensible a factores como la iluminación, el fondo que aparece tras la cara, etc. Por ello se suelen emplear características más elaboradas, como por ejemplo las consideradas por FaceNet (Schroff et al., 2015), método utilizado en el sistema. Consiste en una técnica de *Deep Learning* basada en Redes Neuronales Convolucionales que extrae un vector de 128 dimensiones codificando características faciales como la forma de la cara, la ubicación de ojos, nariz y boca, textura de la piel, etc. Este descriptor presenta cierta invarianza ante cambios como la iluminación, la pose facial y la expresión facial.

4.4. Clasificador

Codificadas las caras como vectores de características, el paso restante para identificar a la persona consiste en clasificar cada cara como perteneciente a una persona conocida o desconocida. Para el objetivo de este trabajo, no es relevante la identidad de la persona, solo que el sistema sea capaz de identificarla para orientarse hacia ella incluso en escenarios con múltiples personas.

De manera simplificada, para clasificar una cara dado su vector de características f_i, este se compara con el vector de características del resto de personas conocidas $f_j, j \in [1, N_p]$ empleando la distancia coseno: $cos\theta = \frac{f_i \cdot f_j}{|f_i||f_j|}$. De este modo, se busca la mayor distancia coseno y, si esta es mayor que un cierto umbral, se clasifica como perteneciente a dicha persona. Además, se emplea su vector de características para actualizar el vector almacenado (entrenamiento en tiempo real). En caso de no superar el umbral, se asume que es una cara nueva y se le asigna un identificador único, permitiendo su reconocimiento en imágenes posteriores.

5. Movimiento del robot

El sistema propuesto actúa en dos situaciones sobre la orientación del robot: al identificar el comienzo de una posible interacción, y al detectar la cara del interlocutor. En cualquier caso, el movimiento a realizar por el robot dependerá de sus capacidades. Por ejemplo, un robot que cuente con los sensores (cámara y micrófonos) montados sobre su carcasa formando un sólido rígido, deberá emplear sus ruedas para rotar y orientarse hacia la persona. Por contra, si cuenta con una cabeza articulada sobre la que se colocan los sensores, podrá rotar dicha cabeza, o incluso combinar la rotación del cuerpo y cabeza. Esta sección describe los movimientos a realizar para orientar la parte en la que se encuentren los sensores.

Así, una vez se ha detectado el comienzo de una posible interacción, el robot comienza un movimiento de rotación para orientarse de manera aproximada hacia su interlocutor, empleando para ello el ángulo de incidencia del sonido estimado (recuérdese la Sec.3.3). Una vez el sistema detecta la cara de su interlocutor y entra en el modo interactuar, se establece un

bucle de control para orientar al robot. En dicho bucle, la consigna es que la cara esté situada en el centro de la imagen, enviando los comandos de movimiento necesarios al robot para que esto sea así, y utilizando nuevas detecciones de la cara para cerrar el bucle. En concreto, el error sobre el que se actúa es la diferencia entre el vector de proyección de la cara detectada con respecto al vector correspondiente cuando esta se encuentra centrada. En situaciones donde el robot sólo pueda orientarse horizontalmente, se actuará sobre el ángulo α_x (rotación en eje X), mientras que si también puede hacerlo verticalmente se incluirá α_y (rotación en eje Y).

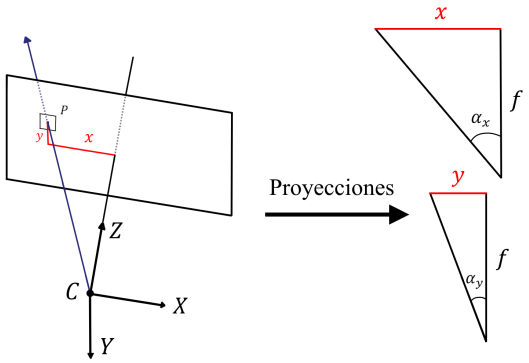

Figura 5: Cálculo de ángulos de rotación respecto a un punto en el espacio

Como se puede apreciar en la Fig. 5, los ángulos permanecen invariantes a lo largo del vector de proyección de la cara. Estos ángulos indican el ángulo que necesita rotar la cámara para centrar su eje óptico en el punto medio del cuadrado que contiene el rostro, siendo independiente de la profundidad a la que se encuentra la persona. Conocidos la matriz de calibración K de la cámara y el punto P del centro del cuadrado que contiene el rostro ($\tilde{P}$ en coordenadas homogéneas), estos ángulos pueden recuperarse como:

$$\begin{bmatrix} x/f \\ y/f \\ 1 \end{bmatrix} = K^{-1}\tilde{P}, \quad \alpha_x = \arctan\frac{x}{f}, \quad \alpha_y = \arctan\frac{y}{f} \qquad (7)$$

De esta forma el bucle de control enviará los comandos necesarios a los actuadores para aproximar α_x y α_y a 0. Ya que la detección facial se realiza sobre el flujo de imágenes procedente de la cámara, esto resulta en un seguimiento constante del interlocutor hasta que la interacción finalice.

6. Validación

Esta sección describe la validación realizada del sistema multimodal propuesto. Concretamente, se describe el robot móvil Sancho (Sec. 6.1), empleado en las pruebas, los detalles de la implementación realizada en ROS2 (Sec. 6.2) y, finalmente, los casos de uso que ejemplifican el correcto funcionamiento del sistema en distintas situaciones (Sec. 6.3).

6.1. *El robot móvil Sancho*

Sancho es un robot móvil de servicio especialmente diseñado para poder interactuar con humanos. Para ello cuenta con una base motorizada y una cabeza móvil articulada (ver Fig. 6). Dicha articulación permite una movilidad completa en dos ejes principales de la cabeza robótica: inclinación (*tilt*) en un rango de aproximadamente 180 grados, y paneo (*pan*) en un rango de movimiento de alrededor de 300 grados. Este sistema es crucial para permitir al robot interactuar de forma efectiva con su entorno, ajustando su campo de visión tanto vertical como horizontalmente. Esto se consigue gracias a la utilización de una unidad *pan-tilt* WidowX XM430 Robot Turret, que incluye los servomotores DYNAMIXEL XM430-W350-T. Esta unidad soporta cargas de más de 2 Kg de peso y permite mover la cabeza con una resolución de 4096 posiciones. Destacar que el sistema proporciona retroalimentación en tiempo real sobre la posición de los ejes, permitiendo la monitorización constante y ajustes dinámicos según las necesidades operativas. En las pruebas realizadas, la cabeza robótica es el elemento que se orienta hacia el interlocutor.

Por parte de su sistema sensorial, este cuenta con numerosos dispositivos para diferentes funcionalidades: escáneres láser para el mapeo, cámara RGB-D, micrófonos, altavoces, etc, aunque únicamente la cámara ojo de pez y los micrófonos montados en su cabeza son relevantes a este trabajo.

Figura 6: Cabeza del robot móvil Sancho

6.2. *Implementación en ROS2*

El sistema ha sido implementado haciendo uso del *framework* de desarrollo de robots ROS2 en su versión *Humble Hawksbill*. La lógica de esta implementación se ha dividido en dos paquetes. El primero de ellos es el encargado del procesamiento del sonido (Sec. 2), el cual incluye nodos responsables de: i) capturar audio, ii) detectar voz humana, y iii) calcular el ángulo de incidencia del sonido. El segundo paquete es el encargado de procesar las imágenes (Sec. 4), incluyendo nodos para: i) capturar imágenes, ii) detectar caras, iii) identificar caras, y iv) gestionar la lógica interna e intercomunicación. Señalar que el funcionamiento de la unidad *pan-tilt* de la cabeza está también integrado en ROS2, permitiendo su configuración para ajustar su posición y su velocidad de movimiento.

Este diseño modular y flexible permite una fácil escalabilidad y mantenimiento del sistema, aprovechando las capacidades de ROS2 para la construcción de aplicaciones robóticas complejas. Además, para el despliegue del sistema se utilizó una arquitectura distribuida basada en contenedores Linux (Ambrosio-Cestero et al., 2024), lo que simplificó el manejo de dependencias y permitió un aprovechamiento eficiente del *hardware* mediante la utilización de recursos en el borde.

6.3. *Casos de uso*

El sistema se ha validado en escenarios límite, donde su naturaleza multimodal le permite determinar con éxito la orientación. La Fig. 7 muestra, para los tres casos de uso descritos, el escenario tal y como lo percibía la cámara en el inicio

de la posible interacción (columna $t_{inicial}$), y su percepción una vez completada la orientación del robot (columna t_{final}).

Caso 1: Cara en FoV + seguimiento.

Ilustrado en la primera fila de la Fig. 7, este sería el escenario más simple: sólo existe un interlocutor y su cara aparece en el campo de visión (del inglés *Field of View*, FoV). Al detectarse tanto el inicio de la interacción como la cara, el sistema rota la cabeza convenientemente para mirar a la persona. La no utilización del módulo de detección de interacción requeriría técnicas adicionales para su detección empleando imágenes.

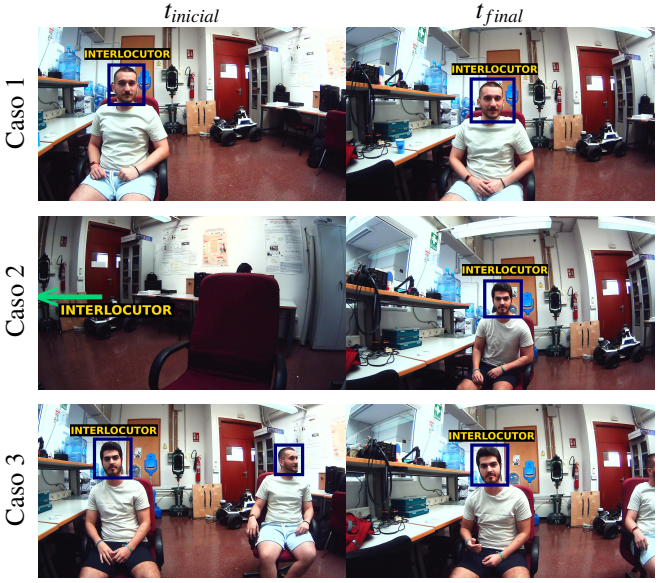

Figura 7: Escenario observado por la cámara al detectarse la interacción ($t_{inicial}$) y al completarse la orientación (t_{final}) en los distintos casos de uso. Caso 1: seguimiento de única cara en FoV. Caso 2: seguimiento de única cara fuera de FoV. Caso 3: Seguimiento ante múltiples caras.

Caso 2: Cara fuera de FoV + seguimiento.

Este caso, mostrado en la segunda fila de la Fig. 7, añade complejidad al anterior al encontrarse la persona fuera del FoV de la cámara. De este modo, el módulo de detección de orientación se encarga de detectar la voz, y de comandar a la unidad *pan-tilt* para que realice una rotación horizontal aproximada en función del ángulo α estimado. Por su parte, el módulo de detección facial permite identificar al interlocutor y realizar una orientación precisa, también en el eje vertical, y un seguimiento activo.

Caso 3: Varias caras en FoV + seguimiento de conversación.

El último caso de uso, ilustrado en la última fila de la Fig. 7, propone un escenario donde de nuevo la multimodalidad propuesta es necesaria para una correcta orientación. En él aparecen múltiples personas en el FoV de la cámara. Así, el ángulo de incidencia del sonido asiste al detector facial para seleccionar de manera coherente a la persona correcta que está asumiendo el rol de interlocutor, de manera similar a como nos comportamos los humanos. La identificación del interlocutor también permite realizar un seguimiento robusto.

7. Conclusiones y trabajos futuros

Este trabajo ha presentado un sistema multimodal para la orientación de un robot móvil hacia su interlocutor en escenarios de interacción humano-robot. La multimodalidad radica en la utilización de **sonido** para la detección del inicio/fin de la interacción, así como para la estimación aproximada del ángulo de incidencia del sonido, lo que permite iniciar una rotación para orientarse, y el uso de **visión** para la localización de la persona empleando detección facial en imágenes, permitiendo una orientación más precisa. La propuesta, implementada en ROS2, ha sido validada con el robot de servicio Sancho, mostrando distintos casos de uso donde la multimodalidad es clave para conseguir una orientación satisfactoria.

Como trabajo futuro, se plantea modificar la configuración actual de micrófonos estéreo añadiendo un tercer micrófono. Esto permitiría superar las limitaciones del sistema actual, ya que posibilitaría discernir si el sonido proviene de la parte delantera o trasera del robot, y estimar el ángulo vertical de incidencia, pudiendo así el robot orientarse hacia personas que se encuentren fuera del FoV en el eje vertical.

Agradecimientos

Este trabajo ha sido desarrollado en el contexto de los proyectos ARPEGGIO (PID2020-117057GB-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, J.-L., Ruiz-Sarmiento, J.-R., Gonzalez-Jimenez, J., 2024. Container based architecture for mobile robotics. XLV Jornadas de Automática.

Baltanas-Molero, S.-F., Ruiz-Sarmiento, J. R., Gonzalez-Jimenez, J., 2020. A face recognition system for assistive robots. In: International Conference on Applications of Intelligent Systems (APPIS).
DOI: https://doi.org/10.1145/3378184.3378225

Baltanas-Molero, S.-F., Ruiz-Sarmiento, J. R., Gonzalez-Jimenez, J., jan 2021. Improving the head pose variation problem in face recognition for mobile robots. Sensors 21 (2).
DOI: https://doi.org/10.3390/s21020659

Bredin, H., Laurent, A., August 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In: Proc. Interspeech 2021. Brno, Czech Republic.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., Gill, M.-P., May 2020. pyannote.audio: neural building blocks for speaker diarization. In: ICASSP 2020.
DOI: https://doi.org/10.1109/ICASSP40776.2020.9052974

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

Macenski, S., Foote, T., Gerkey, B., Lalancette, C., Woodall, W., 2022. Robot operating system 2: Design, architecture, and uses in the wild. Science Robotics 7 (66), eabm6074.
DOI: https://doi.org/10.1126/scirobotics.abm6074

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

Ruiz-Sarmiento, J., Galindo, C., Gonzalez, J., 2011. Improving human face detection through tof cameras for ambient intelligence applications. In: 2nd International Symposium on Ambient Intelligence. pp. 125–132.
DOI: https://doi.org/10.1007/978-3-642-19937-0_16

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823.

Sheridan, T. B., 2016. Human–robot interaction: status and challenges. Human factors 58 (4), 525–532.
DOI: https://doi.org/10.1177/0018720816644364

Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE Signal Processing Letters 23 (10), 1499–1503.
DOI: https://doi.org/10.1109/LSP.2016.2603342