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

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

Señales que hablan: Percepción visual para describir escenas a partir de gestos deícticos en robótica social

Jesús García-Martínez *, Javier Sevilla-Salcedo, José Carlos Castillo Montoya, Álvaro Castro-González, Miguel Ángel Salichs

Departamento de Ingeniería de Sistemas y Automática, Universidad Carlos III de Madrid. Avenida de la Universidad, 30. 28911 Leganés, Madrid. España.

Resumen

La interacción humano-robot busca establecer una comunicación natural combinando elementos verbales y no verbales, siendo especialmente relevante coordinar la atención entre los agentes implicados, un proceso conocido como atención compartida. Aunque la atención compartida basada en el seguimiento de la mirada ha sido ampliamente explorada, el uso de gestos deícticos como mecanismo para guiar la atención ha sido poco abordado en el contexto de la interacción. Este artículo presenta una aplicación interactiva integrada en el robot social Mini, combinando nuestro método previo basado en visión por computador RGB-D para detectar donde señalan los usuarios con modelos generativos multimodales de visión y lenguaje. Nuestra propuesta utiliza la región señalada por el usuario como entrada directa al modelo, generando descripciones verbales coherentes y contextualizadas sobre dicha región. El sistema estima dicha región proyectando un cono tridimensional a partir del brazo del usuario sobre la nube de puntos capturada por el robot, identificando el punto de intersección como foco de atención y definiendo en torno a él una región de interés. Los resultados muestran que el sistema permite al robot generar descripciones precisas y relevantes sobre la zona indicada, mejorando la fluidez y coherencia de la interacción.

Palabras clave: Interacción humano-robot, Robots sociales, Percepción visual, Sistemas multimodales, Atención Compartida, Reconocimiento de gestos, Sistemas inteligentes integrados, Aplicaciones interactivas, Modelos lenguaje-visual

When Gestures Speaks: Visual Perception for Scene Description from Deictic Gestures in Social Robotics

Abstract

Human-robot interaction aims to establish natural communication by combining verbal and non-verbal elements, and it is particularly relevant to coordinate attention between the agents involved, a process known as joint attention. Although joint attention based on gaze tracking has been widely explored, the use of deictic gestures to guide attention has been little addressed in interaction. This paper presents an interactive application embedded in the Mini social robot, combining our previous method based on RGB-D computer vision to detect where users point with multimodal generative models of vision and language. Our approach uses the region pointed to by the user as direct input to the model, generating coherent and contextualised verbal descriptions of that region. The system estimates the region by projecting a three-dimensional cone from the user's arm onto the point cloud captured by the robot, identifying the intersection point as the focus of attention and defining a region of interest around it. The results show that the system allows the robot to generate accurate and relevant descriptions of the indicated area, improving the fluidity and coherence of the interaction.

Keywords: Human-Robot Interaction, Social Robots, Visual Perception, Multimodal Systems, Joint Attention, Gesture Recognition, Embedded Intelligent Systems, Interactive Applications, Vision-Language Models

*Jesús García-Martínez: jesusgar@ing.uc3m.es

Correos electrónicos: jesusgar@ing.uc3m.es (Jesús García-Martínez), javier.sevilla@uc3m.es (Javier Sevilla-Salcedo), jocastil@ing.uc3m.es (José Carlos Castillo Montoya), acgonzal@ing.uc3m.es (Álvaro Castro-González), salichs@ing.uc3m.es (Miguel Ángel Salichs)

1. Introducción

La interacción humano-robot (HRI, por sus siglas en inglés) se refiere al campo de estudio que analiza, diseña y evalúa las formas en que los humanos y los robots se comunican, colaboran y coexisten. Implica tanto aspectos técnicos (como el diseño del robot y su capacidad para percibir e interpretar la conducta humana) como aspectos sociales y psicológicos (como la confianza, la aceptación o la ética en el uso de robots) (Goodrich et al., 2008). De acuerdo con Bonarini (2020), la HRI no solo está limitada a la interacción verbal sino que ha de tener en cuenta otros aspectos definidos como señales sociales, como por ejemplo los gestos, expresiones faciales y lenguaje no verbal que realizamos durante la interacción.

En este contexto, la atención compartida (AC) hace referencia a la capacidad para compartir intencionadamente el interés hacia un objeto, evento, o región específica del entorno mediante señales comunicativas tales como la mirada, el lenguaje o los gestos corporales (Mundy and Newell, 2007; Moore et al., 2014). La comunicación no verbal tiene gran importancia en las interacciones humanas, ya que complementa y enriquece el significado de las palabras, proporcionando "pistas" para interpretar las intenciones y emociones subyacentes del interlocutor (Knapp et al., 1972). La AC facilita que se consiga el entendimiento intersubjetivo, es decir, la construcción recíproca y compartida del significado durante la interacción (Trevarthen and Aitken, 2001; Gallagher and Hutto, 2008). Aplicar estos mecanismos en la robótica social contribuye a que la interacción se perciba como más natural y fluida, al permitir que el robot interprete adecuadamente las señales no verbales del usuario y actúe en coherencia con el contexto. Se ha demostrado que la incorporación de mecanismos de AC mejora significativamente la percepción del usuario en términos de competencia y calidez del robot (García-Martínez et al., 2024), así como su presencia social, entendida como la sensación de que el robot está realmente consciente y comprometido durante la interacción (García-Martínez et al., 2025).

Aunque gran parte de los estudios sobre AC en robótica social han centrado sus esfuerzos en la interpretación de señales basadas en el seguimiento de la mirada, dada su relevancia en la comunicación no verbal humana (Saran et al., 2018; Skantze et al., 2014), otras señales, como los gestos deícticos, aún permanecen poco exploradas. Estos gestos se usan principalmente con la mano, pero también con otras partes del cuerpo como la cabeza o los ojos, con el objetivo de dirigir la atención del receptor hacia un objeto específico. En un trabajo previo, presentamos un algoritmo basado en visión por computador que permite estimar la región de atención en una imagen captada por la cámara de un robot social a partir de la zona donde señala un usuario en un entorno real (Martínez et al., 2024). El método utiliza información RGB-D para detectar los brazos del usuario y estimar la región señalada mediante la intersección del gesto con los objetos de la escena, permitiendo al robot identificar el foco de atención (estímulo más relevante) y redirigir su mirada.

Para interpretar correctamente la acción de señalar no basta con identificar hacia dónde apunta el usuario, sino que también es importante el contexto comunicativo en el que ocurre dicha acción. Actualmente, los modelos multimodales de visión-lenguaje (VLM, por sus siglas en inglés), como CLIP (Radford et al., 2021) o BLIP (Li et al., 2022), han revolucionado la capacidad de describir escenas visuales mediante lenguaje natural. Estos modelos pueden proporcionar explicaciones, descripciones detalladas o incluso valoraciones subjetivas sobre el contenido visual. No obstante, suelen analizar toda una imagen o emplean modelos auxiliares como SAM (Kirillov et al., 2023) para destacar regiones específicas, basados únicamente en *prompts* textuales. Hasta donde alcanza nuestro conocimiento, no existe ningún sistema que utilice las señales manuales del usuario como entrada directa para definir la región que el modelo debe describir en un contexto de HRI.

La principal contribución de este artículo es una aplicación interactiva, implementada sobre el robot social Mini (Salichs et al., 2020), que combina nuestro algoritmo de percepción para la estimación dinámica de la región de atención (foco de atención) con un modelo VLM para la descripción de la misma. Esta integración permite al robot interpretar las señales manuales del usuario en tiempo real, delimitando la región de interés (ROI, por sus siglas en inglés) señalada, y generando verbalmente respuestas adaptadas al contexto interactivo. El robot no solo es capaz de describir visualmente la región señalada, sino también de proporcionar explicaciones contextuales o valoraciones subjetivas, dependiendo del tipo de interacción.

El artículo se estructura de la siguiente manera: la Sección 2 detalla las herramientas y plataformas utilizadas, incluyendo una descripción breve del robot social Mini y el modelo VLM empleado. En la Sección 3, se presenta un caso práctico donde se explica las diferentes etapas por las que pasa la aplicación incluyendo el algoritmo de percepción, junto con la integración del módulo de generación adaptativa de lenguaje natural mediante un ejemplo ilustrativo claro. Finalmente, en la Sección 4 se presentan las conclusiones principales y se indican posibles líneas futuras de investigación.

2. Materiales y Métodos

En esta sección se describen los materiales y métodos empleados para el desarrollo e integración de la aplicación para la descripciones de escenas a través de gestos deícticos en el robot social Mini. En primer lugar, se presenta la plataforma robótica utilizada, detallando sus capacidades de percepción, procesamiento y actuación. A continuación, se analizan distintos modelos generativos multimodales actuales, evaluando sus características y potencial en el contexto de la HRI, y se justifica la selección del modelo finalmente integrado en el sistema.

2.1. Robot Social: Mini

Mini es un robot de sobremesa cuyo objetivo principal es brindar apoyo, compañía y entretenimiento a personas mayores (Salichs et al., 2020). Actualmente cuenta con diversas aplicaciones interactivas y juegos que permiten, entre otras funcionalidades, mostrar imágenes y vídeos, relatar noticias, así como proporcionar actividades recreativas.

Mini tiene una apariencia humanoide con una cubierta exterior de peluche (ver Figura 1). El robot cuenta con cinco grados de libertad distribuidos entre la cabeza, el cuello, los brazos y el torso, lo que le permite realizar movimientos expresivos y coordinados. Está equipado con micrófonos para captar la voz del usuario, altavoces integrados para la generación de habla,

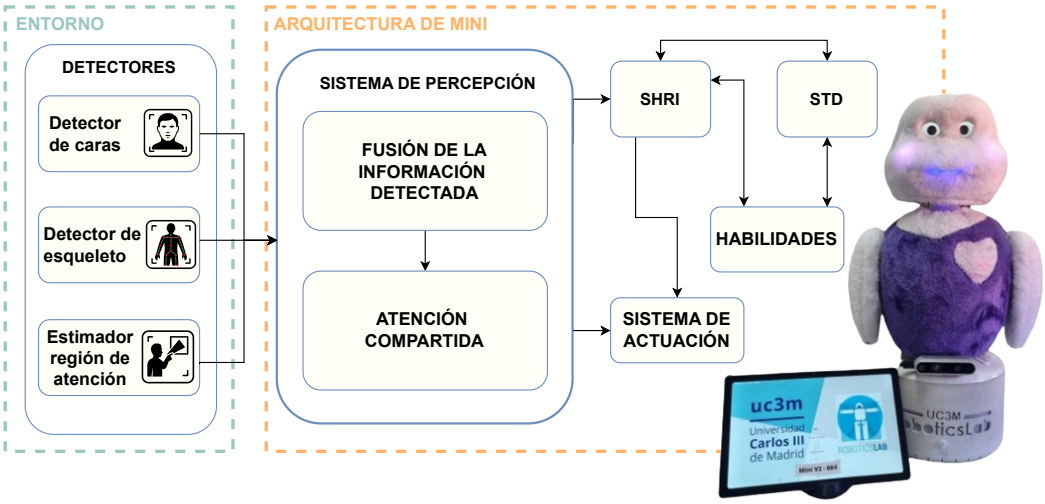


Figura 1: Representación esquemática de la arquitectura software del robot Mini.

sensores táctiles en los brazos y el torso, iluminación LED en las mejillas y boca, y pantallas LED que simulan ojos animados capaces de reflejar distintos estados emocionales. Además, Mini dispone de una tableta auxiliar que amplía sus capacidades comunicativas, permitiéndole mostrar imágenes, vídeos y contenido interactivo, lo que refuerza su carácter multimodal y facilita una interacción natural utilizando canales visuales, auditivos y táctiles.

En términos de percepción, Mini cuenta con cámaras RGB y RGB-D (Intel RealSense D435i[1]) instaladas en la zona de la barbilla y torso. Esto le permite captar información relevante no solo del usuario, sino también del entorno en el que se encuentra, facilitando respuestas apropiadas y contextualizadas. Además, el robot incorpora aceleradores gráficos específicos como Google Coral TPU[2] e Intel Movidius NCS2[3], lo que posibilita la ejecución en local de modelos basados en inteligencia artificial.

Desde el punto de vista de la arquitectura software, Mini está construido sobre ROS (*Robot Operating System*) (Quigley et al., 2009), contando con cinco módulos principales ilustrados en la Figura 1. Dichos módulos son los encargados de gestionar y llevar a cabo las diferentes actividades del robot:

- El **sistema de percepción** es responsable de la adquisición y procesamiento de información procedente del entorno y del usuario con el que Mini interactúa, permitiendo al sistema organizar los diferentes estímulos y gestionar su importancia. En particular, en este trabajo se emplea el detector facial, el detector de esqueletos y el detector propuesto para la estimación de la región de atención.
- El **sistema de toma de decisiones (STD)** ejecuta y gestiona las diferentes aplicaciones (habilidades) que debe realizar el robot en cada momento.
- Las **habilidades** son las diferentes aplicaciones de las que dispone el robot, como los juegos de estimulación y entretenimiento, la reproducción de contenido multimedia, o, en particular, la aplicación de la descripción de escenas propuesta en el presente trabajo.
- El **sistema de interacción humano-robot (SHRI)** procesa la información recibida del sistema de percepción y genera, en función del contexto de la interacción, respuestas multimodales. Estas respuestas incluyen tanto expresiones verbales, mediante síntesis de voz, como no verbales a través de gestos predefinidos. Adicionalmente, el SHRI realiza llamadas a un servidor externo dedicado para ejecutar las inferencias a el modelo multimodal generativo utilizado en este trabajo.
- Por último, el **sistema de actuación** se encarga del control específico de los elementos físicos del robot, incluyendo el control de los motores y la iluminación LED.

2.2. *Modelos multimodales generativos para la descripción de escenas*

En los últimos años, los modelos generativos multimodales han experimentado notables avances gracias a arquitecturas basadas en Transformers capaces de integrar múltiples modalidades sensoriales (imagen, texto, audio). Destaca GPT-4, desarrollado por OpenAI, por su capacidad para combinar entradas visuales y lingüísticas y generar descripciones contextuales precisas, lo que resulta especialmente relevante para tareas comunicativas humano-robot (OpenAI and et al., 2024).

Además de GPT-4, este trabajo ha considerado modelos recientes y relevantes en la literatura, tales como Kosmos-2 (Peng et al., 2023), Flamingo (Alayrac et al., 2022) y CLIP (Radford et al., 2021). Kosmos-2 integra información visual y textual mediante una arquitectura autoregresiva y entrenamiento multimodal a gran escala; Flamingo combina módulos especializados en visión y lenguaje y destaca por sus capacidades en *few-shot*

[1] https://intelrealsense.com/depth-camera-d435i/
[2] https://coral.ai/products/accelerator
[3] https://intel.com/content/www/us/en/products/sku/140109/intel-neural-compute-stick-2/specifications.html

learning; por su parte, CLIP, basado en aprendizaje contrastivo, destaca en clasificación y emparejamiento semántico entre imágenes y textos, aunque no es estrictamente generativo.

Para fundamentar la selección de modelo principal, se han considerado los principales benchmarks reportados en el *OpenVLM Leaderboard* (Duan et al., 2024). En dichas métricas, GPT-4 muestra un rendimiento agregado netamente superior (*Avg Score*=75.9; *Avg Rank*=22.1) frente a Kosmos-2, Flamingo y variantes basadas en CLIP, cuyos resultados se sitúan considerablemente por debajo. Por tanto, GPT-4 ha sido finalmente adoptado como columna vertebral de nuestro sistema de integración visual-lingüística; la integración concreta se describe en secciones posteriores.

3. Aplicación para la descripción de escenas mediante señalización manual

Esta sección describe, a través de un ejemplo representativo, el funcionamiento completo del sistema propuesto, detallando cada una de las etapas implicadas en la interacción entre el usuario y el robot, desde la percepción inicial hasta la generación de la respuesta verbal. En el escenario considerado, el usuario interactúa con el robot en un entorno conocido (p.ej., un laboratorio) y, en un momento dado, realiza un gesto manual señalando un objeto presente en la escena (p.ej. una fotografía colgada en la pared). El objetivo es que el robot interprete correctamente la señal manual, identifique la región de atención indicada y genere una respuesta verbal coherente y adaptada al contexto. La Figura 2 ilustra este ejemplo, que servirá como hilo conductor para el resto de la sección.

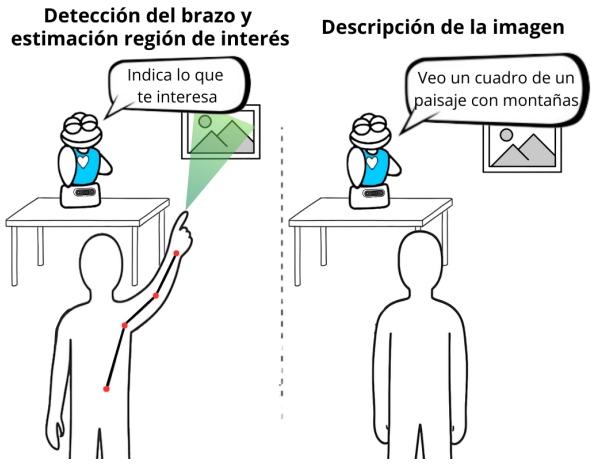
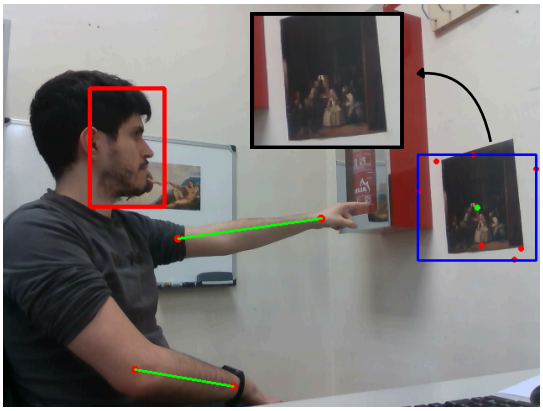


Figura 2: Ejemplo gráfico del proceso propuesto, mostrando la detección del brazo y estimación de la región de interés (izquierda), seguida por la descripción verbal del objeto identificado (derecha).

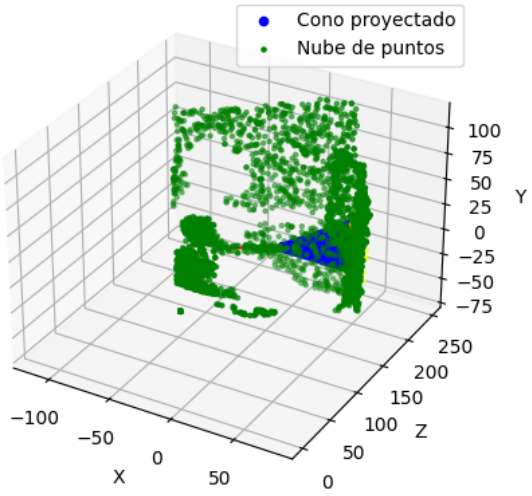
3.1. *Detección del usuario y filtrado de los brazos*

La interacción con el usuario comienza una vez detectada su presencia frente al robot. Para ello, se emplea el modelo FaceNet (Schroff et al., 2015), utilizado para el reconocimiento facial y la localización del usuario dentro de la escena. La salida de este modelo se representa mediante un rectángulo delimitador que encuadra el rostro del usuario en la imagen (ver Figura 3a, rectángulo rojo). Una vez identificado, Mini inicia la interacción y solicita al usuario que señale el objeto o área de interés, manteniendo el brazo extendido durante unos segundos para facilitar la captura visual por parte del sistema.

Para detectar y aislar el brazo utilizado se emplea PoseNet (Kendall et al., 2015), un modelo de visión artificial especializado en la estimación de posturas humanas. PoseNet localiza 17 puntos característicos, entre los cuales destacan codos y muñecas, que son específicamente seleccionados según el brazo utilizado por el usuario (véase Figura 3a, líneas verdes con puntos rojos para el brazo). Posteriormente, las coordenadas bidimensionales de estos puntos relevantes se transforman en coordenadas tridimensionales mediante la proyección sobre el mapa de profundidad proporcionado por la cámara RGB-D, haciendo uso de las funciones de proyección incluidas en la librería oficial de Intel RealSense. Una vez obtenidos dichos puntos 3D, se calcula la zona a la que apunta el usuario como se describe a continuación.



(a) Detección del brazo, rostro, cálculo y recorte de la ROI.



(b) Representación de la nube de puntos tridimensional y del cono virtual.

Figura 3: Ejemplo del proceso para detectar y señalar una ROI mediante la dirección del brazo del usuario captado por Mini. La subfigura 3a muestra la detección del brazo, rostro, ROI estimada e intersecciones con la nube de puntos. La subfigura 3b presenta la nube del entorno y el cono virtual empleado para estimar esta región.

3.2. *Cálculo de proyección*

Utilizando la información del brazo detectado (segmento codo-muñeca), Mini lleva a cabo los siguientes pasos: (1) Construcción de una recta alineada con la dirección del brazo, que se interpreta como la trayectoria del gesto de apuntado. (2) Cálculo de la intersección de dicha recta con la nube de puntos tridimensional de la escena previamente muestreada. Esta intersección aparece como un punto verde en la Figura 3a. (3) Creación de un cono virtual de amplitud variable cuyos vértices corresponden con los extremos del volumen entre la muñeca y el punto intersección (Figura 3b, puntos azules). (4) Filtrado de puntos de la nube original cercanos a la superficie y base del cono (marcados en rojo en Figura 3a interior del ROI azul) que son los utilizados para el cálculo de la región de atención.

3.3. *Estimación de la región de interés*

Partiendo de los puntos tridimensionales del subconjunto filtrado, se lleva a cabo la proyección inversa de los mismos en el plano de la imagen RGB capturada originalmente (dentro del rectángulo azul de la Figura 3a). El proceso consta de tres etapas principales: En primer lugar se realiza la conversión de las coordenadas tridimensionales del subconjunto a coordenadas bidimensionales en el espacio de la imagen, lo que permite ubicar visualmente los puntos sobre el plano de la cámara. A continuación, se calcula el rectángulo delimitante más pequeño (ROI inicial, que engloba todos los puntos proyectados, definiendo una primera aproximación a la región de atención. Finalmente, esta región se refina mediante la aplicación de operadores morfológicos, detección de contornos y técnicas de crecimiento de regiones para mejorar la precisión de dicha región estimada. La región de atención estimada en la imagen 2D corresponde a la entrada empleada posteriormente por el modelo VLM, que será el responsable de generar una descripción textual coherente con el objeto señalado.

3.4. *Descripción de imagen con VLM*

La última etapa del proceso consiste en generar una descripción lingüística para la ROI previamente estimada, empleando el modelo generativo multimodal GPT-4 (cuyas características fueron presentadas en la Sección 2). En este paso, la ROI extraída de la imagen completa (ver Figura 3a, ROI borde negro) constituye la entrada visual inicial para el modelo. Concretamente, la imagen correspondiente a la región señalizada es codificada mediante el módulo visual integrado en el modelo multimodal, generando una representación latente que captura tanto los elementos visuales presentes como sus relaciones semánticas e información espacial relevante. Seguidamente, dicha representación latente junto con un *prompt* predefinido para producir automáticamente una descripción verbal en lenguaje natural es utilizada por el módulo generador autorregresivo basado en Transformers, el resultado es una salida la cual describe el contenido de la ROI en forma comprensible y contextualizada para la HRI. Finalmente, Mini comunica al usuario verbalmente el texto generado, proporcionando así una respuesta interactiva fluida y adecuada al contexto comunicativo planteado. A continuación se muestra el prompt utilizado y un ejemplo de las salida del modelo multimodal para la ROI extraída.

Configuración Modelo 1: Prompt utilizado

"Eres un robot social para ancianos, utilizas un lenguaje amigable, poco cargado y cercano. Cuando te muestran imágenes las ves con tus propios ojos, puedes hacer referencia a tu vista. Te estoy señalando algo, identifica y describe lo que ves. No menciones mi mano (si apareciese en la imagen), ni 'imagen', ni comentes la calidad de la misma."

Salida Modelo 1: Las Meninas

"Parece ser "Las Meninas" de Diego Velázquez. En el centro, se encuentra la infanta Margarita rodeada por sus damas de compañía, con el propio Velázquez a la izquierda pintando en un gran lienzo. En el fondo, se puede ver un espejo reflejando a los reyes. Es una obra maestra del arte barroco, conocida por su complejidad y juego de perspectivas."

De esta manera, se completa el proceso general del sistema a través del ejemplo de ejecución descrito, mostrando cómo las técnicas propuestas permiten que el robot interprete eficazmente el gesto del usuario, extraiga información espacial precisa sobre la región señalada, y genere una descripción verbal coherente mediante la integración del modelo multimodal generativo. En el siguiente enlace se puede visualizar una demostración en vídeo de la ejecución de la aplicación, donde se muestra el funcionamiento del sistema integrado en un escenario de interacción real: https://youtu.be/vlvA8Ft72iQ.

4. Conclusiones

En este trabajo se ha presentado una aplicación interactiva orientada a la descripción automática de escenas señaladas por los usuarios en el contexto específico de la robótica social. La propuesta se basa en la continuación y ampliación de una línea previa de trabajo, en la que se emplean algoritmos de visión por computador para identificar e interpretar gestos manuales realizados por un usuario con modelos generativos multimodales capaces de generar respuestas lingüísticas adecuadas a partir del contenido visual señalado. En concreto, se ha desarrollado un sistema interactivo que permite al robot social identificar dinámicamente regiones de atención señaladas mediante gestos deícticos, estimar el área concreta de interés en la escena visual, y proporcionar al usuario una descripción verbal coherente, precisa y contextualizada sobre dicha área.

Desde una perspectiva de HRI, este trabajo contribuye hacia la integración efectiva del lenguaje no verbal, específicamente mediante la consideración explícita de elementos como los gestos deícticos como un canal comunicativo adicional. Al capturar, interpretar e incorporar dicha información contextual no verbal, el sistema favorece un estilo comunicativo más natural, enriquecido, intuitivo y cercano a las interacciones cotidianas. Esto permite ofrecer a los usuarios experiencias considerablemente más familiares, disminuyendo potenciales dificultades en la comunicación, promoviendo la adaptación del usuario

al robot y fortaleciendo el binomio usuario-robot como un equipo comunicativo efectivo, especialmente en contextos como la robótica asistencial o educativa.

La incorporación de modelos generativos multimodales presenta ventajas adicionales frente a aproximaciones más tradicionales o basadas exclusivamente en técnicas clásicas de visión artificial. Por un lado, los VLM aportan mayor flexibilidad a la hora de describir una escena, adaptándose a contextos específicos mediante instrucciones textuales sencillas (*prompting*). Por otro lado, permite ir más allá de descripciones meramente neutras u objetivas, abriendo nuevas vías hacia interacciones más expresivas, que contemplen explicar relaciones visuales más complejas o generar valoraciones subjetivas, opiniones o comparaciones acerca de objetos o situaciones señaladas. Esto puede traducirse en una HRI mucho más rica desde el punto de vista tanto emocional como informativo.

Agradecimientos

Estos resultados han sido financiados por los proyectos PID2021-123941OA-I00, financiado por MCIN/AEI/10.13039/501100011033 y por ERDF A way of making Europe; TED2021-132079B-I00 financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR; Mejora del nivel de madurez tecnológica del robot Mini (MeNiR) financiado por MCIN/AEI/10. 13039/501100011033. 13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR; Robot social portable con alto grado de vinculación (PoSoRo) PID2022-140345OB-I00 financiado por MCIN/AEI/10.13039/501100011033 y ERDF A way of making Europe y por iRoboCity2030-CM, Robótica inteligente para ciudades sostenibles (TEC-2024/TEC-62), funded by Programas de Actividades I+D en tecnologías de la Comunidad de Madrid.

Referencias

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K., 2022. Flamingo: a visual language model for few-shot learning.
URL: `https://arxiv.org/abs/2204.14198`

Bonarini, A., 2020. Communication in human-robot interaction. Current Robotics Reports 1 (4), 279–285.

Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al., 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11198–11201.

Gallagher, S., Hutto, D. D., 2008. Understanding others through primary interaction and narrative practice. In: The shared mind: Perspectives on intersubjectivity. John Benjamins Publishing Company, pp. 17–38.

García-Martínez, J., Gamboa-Montero, J. J., Castillo, J. C., Castro-González, Á., 2024. Analyzing the impact of responding to joint attention on the user perception of the robot in human-robot interaction. Biomimetics 9 (12), 769.

García-Martínez, J., Gamboa-Montero, J. J., Castillo, J. C., Castro-González, Á., Salichs, M. A., 2025. Implementation of a biologically inspired responsive joint attention system for a social robot. Advanced Intelligent Systems, 2400650.

Goodrich, M. A., Schultz, A. C., et al., 2008. Human–robot interaction: a survey. Foundations and trends® in human–computer interaction 1 (3), 203–275.

Kendall, A., Grimes, M., Cipolla, R., 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision. pp. 2938–2946.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026.

Knapp, M. L., Hall, J. A., Horgan, T. G., 1972. Nonverbal communication in human interaction. Thomson Wadsworth.

Li, J., Li, D., Xiong, C., Hoi, S., 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. PMLR, pp. 12888–12900.

Martínez, J. G., Sevilla-Salcedo, J., Montoya, J. C. C., González, Á. C., Sánchez-Caballero, M. Á. S., 2024. Estimando la región de atención mediante atención compartida en robots sociales. In: Actas del Simposio de Robótica, Bioingeniería y Visión por Computador: Badajoz, 29 a 31 de mayo de 2024. Servicio de Publicaciones, pp. 139–144.

Moore, C., Dunham, P. J., Dunham, P., 2014. Joint attention: Its origins and role in development. Psychology Press.

Mundy, P., Newell, L., 2007. Attention, joint attention, and social cognition. Current directions in psychological science 16 (5), 269–274.

OpenAI, et al., 2024. Gpt-4 technical report.
URL: `https://arxiv.org/abs/2303.08774`

Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F., 2023. Kosmos-2: Grounding multimodal large language models to the world.
URL: `https://arxiv.org/abs/2306.14824`

Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A. Y., et al., 2009. Ros: an open-source robot operating system. In: ICRA workshop on open source software. Vol. 3. Kobe, Japan, p. 5.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PmLR, pp. 8748–8763.

Salichs, M. A., Castro-González, Á., Salichs, E., Fernández-Rodicio, E., Maroto-Gómez, M., Gamboa-Montero, J. J., Marques-Villarroya, S., Castillo, J. C., Alonso-Martín, F., Malfaz, M., 2020. Mini: a new social robot for the elderly. International Journal of Social Robotics 12, 1231–1249.

Saran, A., Majumdar, S., Short, E. S., Thomaz, A., Niekum, S., 2018. Human gaze following for human-robot interaction. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 8615–8621.

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.

Skantze, G., Hjalmarsson, A., Oertel, C., 2014. Turn-taking, feedback and joint attention in situated human–robot interaction. Speech Communication 65, 50–66.

Trevarthen, C., Aitken, K. J., 2001. Infant intersubjectivity: Research, theory, and clinical applications. The Journal of Child Psychology and Psychiatry and Allied Disciplines 42 (1), 3–48.