ASISTENTE VISUAL INTERACTIVO PARA DETECCIÓN DE OBJETOS EN SISTEMAS DE VISIÓN PROTÉSICA

Ana Franco-Martinez, Julia Tomas-Barba, Alejandro Perez-Yus

Afiliación: Grupo de Robótica, Visión por computador e Inteligencia Artificial (RoPeRT)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, España.
Tel. +34976762707, e-mail: 821809@unizar.es

Resumen

Este trabajo está enfocado en mejorar la interacción entre una persona invidente con prótesis visual con su entorno. Se ha desarrollado un asistente controlado por voz que recibirá las órdenes de búsqueda utilizando lenguaje natural y ayudará al paciente a localizar objetos en la escena a modo de realidad aumentada.

1. Introducción

Las prótesis visuales son dispositivos asistenciales desarrollados para restaurar parte de la visión a personas que sufren enfermedades degenerativas de la retina. Esto se consigue gracias a estimular eléctricamente las células supervivientes de la retina que generan señales neuronales que son interpretadas como puntos de luz llamados *fosfenos*. Sin embargo, la percepción es limitada debido a diversos factores, como es la baja resolución, el estrecho campo de vista o el ruido, lo que dificulta en gran medida tareas cotidianas tales como buscar las llaves o un lugar para sentarse.

Además, debido al reducido número de pacientes con prótesis visuales y a la dificultad para acceder a ellos, la evaluación de las técnicas propuestas se lleva a cabo a través de la Visión Protésica Simulada (SPV, por sus siglas en inglés). La SPV busca recrear con la mayor fidelidad posible el comportamiento de las prótesis, permitiendo llevar a cabo experimentos con sujetos de visión sana, y de esta manera sacar conclusiones estadísticas.

Existen numerosos trabajos que se centran en la mejora de la calidad de las percepciones visuales generadas por las prótesis. Mientras algunos se centran en la optimización personalizada de los estímulos eléctricos [1], otros abordan tareas de movilidad evitando obstáculos [2].

A diferencia de estas propuestas, este trabajo se centra en la asistencia interactiva mediante un asistente controlado por voz, para facilitar al usuario la detección de objetos. Se integran técnicas de detección visual con sistemas de interacción por voz, mejorando la comprensión del entorno. Finalmente, se usa un SPV realista [1] que permitirá llevar a cabo experimentos controlados con sujetos con visión sana.

2. Asistente visual interactivo

El esquema general del sistema propuesto se presenta en la Figura 1, donde se muestran los tres módulos principales: el asistente de voz, el detector de objetos y el simulador de fosfenos. La figura indica el flujo de información desde la entrada del usuario en lenguaje natural hasta la generación de la percepción visual simulada, incluyendo la fase final de experimentación.

2.1. Asistente de voz

Hemos creado un asistente de voz para facilitar la introducción de las órdenes de búsqueda. Entre los requisitos de diseño buscados, se encuentran la activación con la propia voz, la transcripción de audio a texto, la comprensión del lenguaje natural para reconocer las órdenes de búsqueda, y el funcionamiento en castellano. En concreto, para activar al asistente utilizaremos un sistema de *Wake Word*, que realiza una escucha continua, sin interferir en el resto de tareas que se pueden estar realizando, (de manera similar a, por ejemplo, decir "Alexa" con productos Amazon). Una vez se detecta la palabra clave, el usuario tiene unos segundos para introducir una orden usando lenguaje natural, que se transcribe utilizando, en nuestro caso, la librería *Whisper* de OpenAI.

A partir de la transcripción, se extraen los sustantivos, que tomaremos como los objetos a buscar. Para poder utilizarlos como entrada del detector de objetos (ver Sección 2.2), será necesario traducirlos al inglés, lo cual se lleva a cabo con *Google Translator.* Además, el asistente es capaz de realizar funciones adicionales, como añadir más objetos, borrar alguno, indicar el fin de la búsqueda, o cambiar el modo de visualización.

2.2. Detector de objetos con vocabulario abierto

Una vez identificados los objetos de interés, se emplea un detector de objetos de vocabulario abierto, *OWL-ViT*[3], que es capaz de reconocer categorías que no estaban presentes en los conjuntos de entrenamiento, a diferencia de otros métodos clásicos como *YOLO*. Una característica especialmente relevante para nuestra aplicación, es su capacidad de funcionar en tiempo real. De esta manera, utilizamos las palabras extraídas por el asistente (empleadas como *prompts*) junto a la información visual capturada por la cámara para obtener las *bounding boxes* de los objetos detectados. Para mejorar la interacción del usuario con los objetos, se incorpora un módulo de segmentación que permite estimar el contorno con mayor precisión. Usando las *bounding boxes*, aplicamos *SAM2* [4] para generar una segmentación precisa de los objetos.

2.3. Simulación realista de fosfenos

Tal y como hemos mencionado anteriormente, se utiliza un simulador de visión protésica para poder evaluar la utilidad del método propuesto en unas condiciones similares a las de un usuario real de prótesis visual. Para ello, se emplea el encoder descrito en [1], que optimiza los estímulos eléctricos generados a partir de una imagen de entrada, considerando las características físicas del paciente. A continuación, un modelo computacional simula la percepción visual del usuario generando una representación *fosfénica* realista de la escena en la que se incorporan factores como la resolución espacial limitada o la deformación de los *fosfenos*.

3. Resultados

En esta sección presentamos los resultados cualitativos de nuestro trabajo. En la Figura 2 se muestra, a partir de la imagen original (capturada por la cámara), variantes del proceso de detección de objetos, y sus visualizaciones en la representación fosfénica. Con estos resultados se ilustra la dificultad para detectar objetos si no se lleva a cabo ningún procesamiento, así como justificar la inclusión del segmentador, dado que las *bounding boxes* que proporciona *OWL-ViT* nos muestran una visualización limitada. Finalmente, se presenta el resultado de mostrar la máscara segmentada junto a un oscurecimiento del resto de la escena para así poder resaltar el objeto de interés.

4. Conclusiones

El trabajo que se ha llevado a cabo significa una mejora en cómo el paciente puede relacionarse con su entorno de una manera más fluida y facilita su interacción con la escena. A futuro se espera poder llevar a cabo experimentos en pacientes que porten una simulación de la prótesis utilizando gafas de realidad aumentada para poder probar su eficacia.

REFERENCIAS

[1]. TOMAS-BARBA, Julia, Perez-Yus, A., et al. Adaptive vision transformer for enhanced perception in visual prostheses. IEEE EMBC, 2025

[2]. PEREZ-YUS, Alejandro, et al. RASPV: A robotics framework for augmented simulated prosthetic vision. IEEE Access, 2024, vol. 12, p. 15251-15267.

[3]. MINDERER, Matthias, et al. Simple open-vocabulary object detection. En ECCV, 2022. p. 728-755.

[4]. RAVI, Nikhila, et al. SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024.

Figura 2. Ejemplos con dos objetos ('silla' y 'llaves'), con distintas configuraciones de visualización (sólo detector –OWL– o también segmentador –SAM2– y oscureciendo).

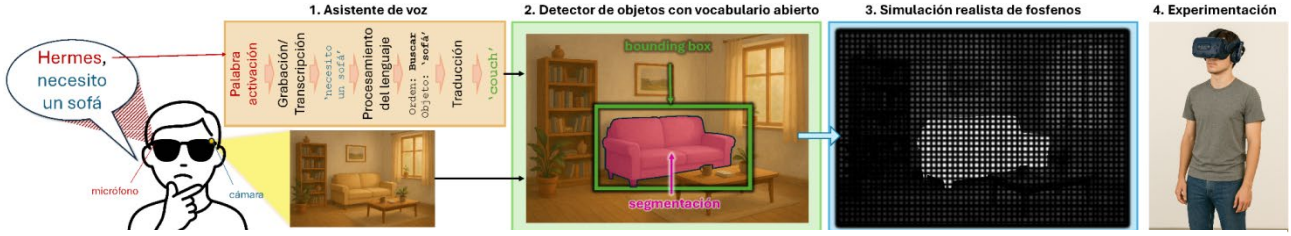

Figura 1. Esquema del asistente visual interactivo desarrollado, que consta de un asistente de voz, un sistema de detección de objetos, y una simulación realista de prótesis visuales. El objetivo final es utilizarlo en experimentación con sujetos reales.