

INTERFAZ VISUAL-AUDITIVA PARA PERSONAS DÉBILES VISUALES

INVESTIGACIÓN

VISUAL-AUDITORY INTERFACE FOR PEOPLE WITH VISUAL IMPAIRMENTS

RESUMEN

La interpretación automática de escenas visuales complejas y dinámicas, como las de un supermercado, representa una frontera significativa para la visión por computadora. La detección robusta de una gran diversidad de objetos y la extracción precisa de información contextual como los precios, en entornos no controlados con iluminación variable y oclusiones, siguen siendo un desafío científico fundamental. Esta limitación impacta directamente el desarrollo de tecnologías asistenciales efectivas para millones de personas con discapacidad visual que enfrentan barreras en tareas cotidianas como las compras autónomas. Este trabajo presenta el desarrollo de un marco computacional novedoso para la interpretación visual-auditiva. La metodología se basa en un modelo de aprendizaje profundo que integra una arquitectura de detección de objetos (YOLOv8, You only look once, versión 8) con un sistema de reconocimiento óptico de caracteres (OCR, por sus siglas en inglés), ambos optimizados para operar eficientemente en un dispositivo móvil con un diseño de lentes. Para entrenar y validar el sistema, se ha creado un nuevo conjunto de datos con miles de imágenes capturadas en supermercados reales, anotadas con cajas delimitadoras para productos y etiquetas de precios. Como resultado principal, nuestro modelo demuestra una mejora significativa en la precisión de detección simultánea de objetos en el contexto de los supermercados en comparación con los enfoques estándar, manteniendo un alto rendimiento en condiciones de baja luminosidad (menores a 450 luxes) y vistas parciales, se alcanzó una confiabilidad de hasta el 80 % en la detección de productos de muestra, manteniendo una latencia operativa mínima. En conclusión, este trabajo no solo ofrece una solución tecnológica viable, sino que también genera conocimiento sobre la adaptación de modelos de inteligencia artificial a entornos no estructurados, sentando las bases para una nueva generación de herramientas asistenciales de bajo costo que mejoran sustancialmente la autonomía e inclusión de las personas con debilidad visual.

Palabras clave: visión por computadora; aprendizaje profundo; debilidad visual; tecnología asistencial.

PÉREZ CASILLAS, Alan Refugio

ZAVALA LUNA, Guadalupe

SANDOVAL GÓMEZ, María Fernanda

GÓMEZ CASAS, Josué

TORO ARCILA, Carlos Alberto

OBREGÓN FLORES, Jonathan
Sebastián

ORTIZ RAMOS, Daniela Estefanía

Facultad de Ingeniería,
Universidad Autónoma de
Coahuila. Arteaga, Coahuila,
México.Correspondencia
jogomezc@uadec.edu.mx
<https://orcid.org/0000-0002-7696-0433>Fecha de recepción
5 de febrero de 2026.Fecha de aceptación
14 de junio de 2026.

ABSTRACT

The automatic interpretation of complex and dynamic visual scenes, such as those found in a supermarket, represents a significant frontier in computer vision. The robust detection of a wide diversity of objects and the precise extraction of contextual information such as prices, in uncontrolled environments with variable lighting and occlusions, remains a fundamental scientific challenge. This limitation directly impacts the development of effective assistive technologies for millions of people with visual impairments who face barriers in everyday tasks such as independent shopping.

This work presents the development of a novel computational framework for visual-auditory interpretation. The methodology is based on a deep learning model that integrates an object detection architecture (YOLOv8) with an optical character recognition (OCR) system, both optimized to operate efficiently on a mobile device with a lens-based design. To train and validate the system, a new dataset was created with thousands of images captured in real supermarkets, annotated with bounding boxes for products and price tags.

As a main result, our model demonstrates a significant improvement in the simultaneous detection accuracy of objects within supermarket contexts compared to standard approaches, while maintaining high performance under low-light conditions and partial views. A reliability of up to 80 % was achieved in the detection of sample products, while maintaining minimal operational latency.

In conclusion, this work not only offers a viable technological solution, but also generates knowledge on adapting artificial intelligence models to unstructured environments, laying the foundation for a new generation of low-cost assistive tools that substantially enhance the autonomy and inclusion of people with visual impairments.

Keywords: computer vision, deep learning; visual impairments; assistive technologies.

INTRODUCCIÓN

La magnitud del desafío de la discapacidad visual, tanto a nivel global como nacional, es apremiante. Según la OMS (2023), más de 2,200 millones de personas viven con deterioro visual, con acceso limitado a soluciones. En México, el INEGI (2020) reporta al menos 2,691,000 personas con deficiencia visual, que representan el 2.14 % de la población mexicana total y enfrentan barreras significativas en su vida diaria. A nivel local (Coahuila y Saltillo), las cifras también son considerables (57,381 y 12,196 personas respectivamente), subrayando la necesidad urgente de intervenciones efectivas. Una de las barreras cotidianas más relevantes es la dificultad para realizar tareas esenciales, como las compras en supermercados, de forma autónoma. Este proyecto aborda directamente esta problemática social al investigar y desarrollar los fundamentos de una herramienta tecnológica asistencial (la interfaz visual-auditiva basada en el marco computacional propuesto). Al permitir la identificación de productos y precios mediante procesamiento local en un dispositivo potencialmente accesible (basado en hardware como la Raspberry Pi 4), se busca ofrecer una solución práctica para mejorar la autonomía y calidad de vida de este sector poblacional.

La búsqueda de soluciones tecnológicas para asistir a personas con debilidad visual en sus actividades cotidianas ha explorado diversas vías. Enfoques iniciales se basaron en tecnologías como RFID (indicadores de radiofrecuencia por sus siglas en inglés) para la identificación de objetos sin contacto, acopladas a sistemas de reproducción de audio para describir el objeto detectado, facilitando tareas básicas (Fernández, 2012). Si bien útiles para objetos pre-etiquetados, estas soluciones presentan limitaciones intrínsecas en entornos dinámicos y no controlados como un supermercado, donde la mayoría de los productos no cuentan con etiquetas RFID y el objetivo es reconocer visualmente una amplia gama de artículos.

Posteriormente, la visión por computadora emergió como un campo prometedor. Propuestas tempranas incluyeron el uso de realidad aumentada en dispositivos tipo lentes ("AR GLASS") para detectar objetos mediante una cámara web y ofrecer retroalimentación auditiva (Castaño Rivera et al., 2015). Aunque conceptualmente interesantes, estos sistemas mostraron dificultades significativas en entornos desconocidos, donde la falta de guía sobre dónde enfocar limitaba su usabilidad práctica. Otros enfoques exploraron la visión estereoscópica para obtener información tridimensional del entorno y traducirla a señales hápticas (vibraciones), orientándose más a la detección

de obstáculos y la navegación que al reconocimiento detallado de objetos (Cruz, 2019). Investigaciones más específicas utilizaron visión artificial para tareas concretas como el reconocimiento de billetes, pero a menudo requerían condiciones controladas (uso de un guante de color específico para delimitar el área de interés) y algoritmos de procesamiento de imagen tradicionales (HSV, binarización) que no escalan bien a la complejidad y variabilidad de escenarios como un supermercado (Tamayo Zapata, 2018). De manera similar, sistemas basados en algoritmos como Haar Cascade se enfocaron en la detección de objetos en entornos interiores, pero con limitaciones en la generalización y robustez ante la diversidad de productos comerciales (Guillen Salazar & Bonilla Bonilla, 2021).

El avance reciente más significativo proviene de la integración de inteligencia artificial (IA) y cámaras en dispositivos portables, como los lentes inteligentes desarrollados por Meta en colaboración con Ray-Ban. La primera generación (Meta, 2021) se centró en la captura multimedia para entretenimiento, con capacidades de procesamiento limitadas para tareas asistenciales complejas (Meta, 2021; Iqbal & Campbell, 2022). La segunda generación (2023) representa un salto cualitativo al incorporar cámaras de mayor resolución y asistentes de IA (Meta AI) capaces, según se informa, de realizar reconocimiento de objetos y traducción visual-auditiva (Waisberg et al., 2023).

Por lo tanto, a pesar de los progresos existentes, persiste una brecha significativa en el conocimiento y la tecnología para desarrollar sistemas de interpretación visual-auditiva que sean verdaderamente robustos, eficientes, específicos para la tarea de compra en supermercados y potencialmente de bajo costo. Este proyecto se enfoca en desarrollar y validar un marco computacional basado en visión por computadora y aprendizaje profundo para la interpretación visual robusta y en tiempo real de productos de abarrotes y sus precios en entornos complejos de supermercado, estableciendo los fundamentos metodológicos para sistemas de asistencia visual-auditiva eficientes y accesibles.

METODOLOGÍA

La presente investigación adopta un enfoque metodológico experimental y cuantitativo, fundamentado en los principios de la visión por computadora, el aprendizaje profundo y el procesamiento de señales. Sigue una serie de pasos específicos mostrados en la figura 1 y se basa en el marco teórico de las redes



neuronales convolucionales (CNN, por sus siglas en inglés), el aprendizaje supervisado y los modelos de percepción artificial, con el objetivo de desarrollar y validar un marco computacional robusto para la interpretación visual en tiempo real de entornos complejos, como los supermercados. La estrategia se enfoca en abordar los desafíos científicos inherentes a esta tarea (variabilidad, densidad, integración objeto-texto) y generar conocimiento sobre la efectividad y limitaciones de los enfoques actuales, culminando en la evaluación de un prototipo para asistencia visual.

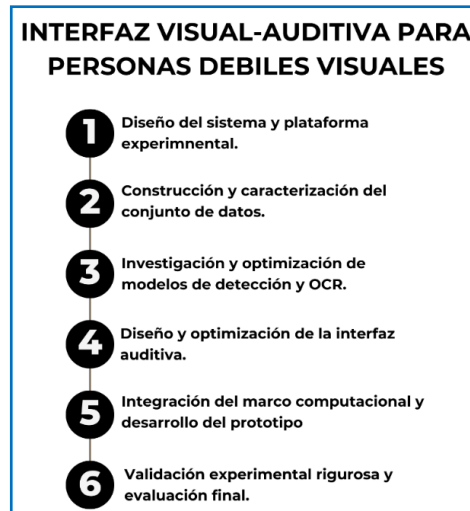


Figura 1. Diagrama de metodología.

DISEÑO DEL SISTEMA Y PLATAFORMA EXPERIMENTAL

Se utilizó una Raspberry Pi 4 modelo B como sistema representativo para la evaluación de rendimiento en el borde (edge computing); la autonomía energética fue provista por una batería portátil recargable de 200000 mAh, completando la configuración de la plataforma experimental para validación. La adquisición de datos visuales se realizó mediante una cámara RGB monocular inalámbrica. La interfaz auditiva se implementará con un módulo de audio, bocinas compactas y/o auriculares inalámbricos.

CONSTRUCCIÓN Y CARACTERIZACIÓN DEL CONJUNTO DE DATOS

Se creó un conjunto de datos específico y representativo del dominio de los supermercados. Esto implicó la recolección sistemática de imágenes utilizando



4737 imágenes de productos de supermercado (categorías definidas: enlatados, cajas de leche, cajas de cereal y papel sanitario), el dataset se dividió en un 80 % entrenamiento y 20 % validación. Se utilizó la plataforma Roboflow para la gestión, anotación precisa (bounding) y versionado del dataset. Se aplicó un conjunto curado de técnicas de Data Augmentation (geométricas: rotación, escalado, traslación; fotométricas: brillo, contraste, saturación; y potencialmente otras como CutMix, Mosaic) específicamente seleccionadas para simular la variabilidad del entorno real y evaluar su impacto en la robustez y generalización de los modelos entrenados.

INVESTIGACIÓN Y OPTIMIZACIÓN DE MODELOS DE DETECCIÓN Y OCR

Se investigó la adaptación del modelo YOLOv8, que adopta un cabezal Ultralytics dividido sin anclaje, lo que contribuye a una mejor precisión y a un proceso de detección más eficiente en comparación con los enfoques basados en anclaje (Ultralytics YOLOv8), para la tarea específica de detección de objetos (productos) y texto pequeño (precios) en nuestro dataset, asimismo se evaluaron y compararon las métricas arrojadas por el reentrenamiento.

DISEÑO Y OPTIMIZACIÓN DE LA INTERFAZ AUDITIVA

Siguiendo los principios de diseño de interfaces auditivas eficaces, se desarrolló el módulo de conversión texto-a-voz utilizando librerías como pyttsx3 o gTTS, evaluando y seleccionando la que ofrezca mejor inteligibilidad y menos latencia en la plataforma Raspberry Pi. Se optimizó el flujo de procesamiento para asegurar que la latencia total desde la captura de imagen hasta la salida auditiva sea consistentemente inferior a 1 segundo, un umbral crítico para la usabilidad en tiempo real. Se considerarán estrategias para presentar la información de forma concisa y no abrumadora.

INTEGRACIÓN DEL MARCO COMPUTACIONAL Y DESARROLLO DEL PROTOTIPO EXPERIMENTAL

Los componentes investigados y optimizados (modelo de detección, estrategia de asociación, interfaz auditiva) se integraron en un marco computacional coherente. Este marco se implementará en la plataforma, constituyendo el prototipo demostrador funcional. Se le realizará una adaptación esencial al modelo para asegurar su funcionamiento en la Raspberry Pi 4.



VALIDACIÓN EXPERIMENTAL RIGUROSA Y EVALUACIÓN FUNCIONAL

Se evaluó la precisión end-to-end (detección+OCR+asociación), latencia y eficiencia en el conjunto de datos de validación final, incluyendo escenarios diseñados específicamente para probar la robustez ante variaciones controladas de iluminación, distancia, ángulo y oclusión.

RESULTADOS Y DISCUSIÓN

Hasta el momento se logró la creación de una red neuronal convolucional capaz de identificar y detectar ciertos productos de supermercado (clases: caja de leche, cajas de cereal, enlatados y papel sanitario). Ésta arrojó unas métricas que muestran alrededor de un 80 % de precisión como se muestra en la figura 2.

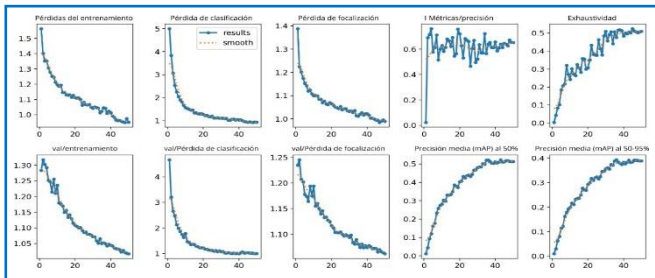


Figura 2. Métricas del reentrenamiento de la red.

Utilizando YOLOv8 comercial, se lograron resultados que arrojaron hasta un 88 % de confiabilidad al momento de la detección e identificación de algunas clases (tazas, botellas, sillas y personas). Estos resultados fueron tomados con una luminosidad de 450 lux a mediodía, mostrados en las figuras 3 y 4.

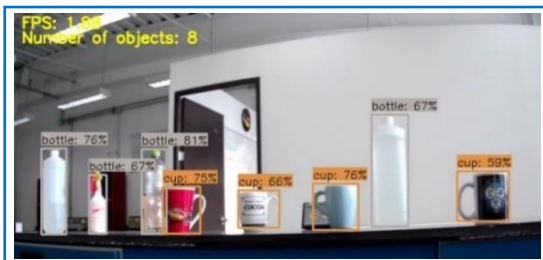


Figura 3 Detección de las clases “tazas y botellas”.



Figura 4. Detección de las clases “personas y sillas”.

Del mismo modo, se obtuvieron resultados favorables con la red reentrenada en la cual se logró la identificación y detección de algunos productos de supermercado (cajas de leche, cajas de cereal, enlatados y papel sanitario), como se observa en las figuras 5 y 6.



Figura 5. Detección de la clase “jugo” con la red reentrenada.

Figura 6. Detección de la clase “caja de cereal” con la red reentrenada.



El resultado científico principal de este proyecto fue el desarrollo y la validación de un marco computacional robusto para la interpretación visual en tiempo real (detección de productos y lectura de precios) en entornos complejos y no controlados, como los supermercados, utilizando visión por computadora y aprendizaje profundo.

En conjunto, estos resultados impulsaron el conocimiento en la intersección de la visión por computadora, el aprendizaje profundo y las tecnologías asistenciales, proporcionando metodologías validadas, caracterizaciones de rendimiento y principios de diseño aplicables a otros problemas de percepción artificial en entornos reales y complejos.

Aunque los resultados fueron favorables, la red se tiene que someter a modificaciones, como el aumento de la calidad de las imágenes en el dataset, para aumentar su precisión (menor del 80 %) para lograr alcanzar o superar el porcentaje de confiabilidad del modelo YOLOv8 comercial (arriba del 80 %). Aunado a esto, se tiene que trabajar con la Raspberry para buscar la mejor manera de adaptación del modelo de red neuronal, teniendo en cuenta las limitaciones del hardware, ya que cuenta con poca memoria RAM, una GPU muy limitada y baja capacidad de procesamiento que ofrece este tipo de plataforma.

CONCLUSIONES

Según los resultados obtenidos hasta el momento, se logró una correcta identificación y detección de objetos utilizando Yolo v8 comercial, con un nivel de confiabilidad de hasta el 88 %. El reentrenamiento del modelo con un mayor número de clases específicas de supermercado resultó en una mejora en la capacidad de detección. Sin embargo, este aumento en la complejidad del problema conllevó una disminución en el porcentaje de confianza promedio, un compromiso conocido en el entrenamiento de redes neuronales, por lo que es necesario trabajar en el reentrenamiento de la red, para que pueda llegar o superar el porcentaje de confiabilidad antes obtenido con Yolo v8 comercial, de la mano de las modificaciones a las cuales se está sometiendo la red, se sigue trabajando con las modificaciones en la Raspberry para que la detección se pueda realizar en tiempo real sin tener un desfase de tiempo tan extenso, ya que hasta el momento se ha logrado trabajar con 2 FPS.

AGRADECIMIENTOS

A la Universidad Autónoma de Coahuila y al Laboratorio de Automatización de la Facultad de Ingeniería, por el apoyo para la realización de este trabajo.

REFERENCIAS

- Castaño, S., Peinado, Y. J. & Carcamona, E. (2013). Coantruccion de un dispositivo de bajo costo como apoyo a personas con discapacidad visual utilizando realidad aumentada, *Ingenieria e Innovacion*, 1(2): 769. DOI: <https://doi.org/10.21897/23460466.769>
- Cruz, S. M. (2019). Desarrollo de prototio para la detección de obstáculos en personas invidentes por medio de visión artificial. Tesis de Maestría en Ciencias con acentuación en Mecatrónica. Universidad Autónoma de Queretaro, Queretaro, Mexico.
- Fernández, L. E. O. (2012). Sistema de identificación de objetos para personas invidentes usando la tecnología RFID. *Ingeniues*, (8): 38-44.
- Guillen Salazar, M. F. & Bonilla Bonilla, T. J. (2021). Diseño e implemetnacion de un prototipo embebido de reconocimiento de objetos por medio de visión artificial como ayuda a personas con deficiencia visual. Tesis de Ingenieria en Electronica y Automatizacion. Escuela Superior Politecnica de Chimborazo, Riobamba, Ecuador.
- INEGI, Instituto Nacional de Estadística y Geografía (2020). Estadística de discapacidad visual. [En línea]. Disponible en: [https://www.gob.mx/profeco/documentos/ojo-con-tu-salud-visual?state=published#:~:text=Datos%20del%20Censo%20de%20Poblaci%C3%B3n,y%20presbicia\)%2C%20catarata%20senil%2C](https://www.gob.mx/profeco/documentos/ojo-con-tu-salud-visual?state=published#:~:text=Datos%20del%20Censo%20de%20Poblaci%C3%B3n,y%20presbicia)%2C%20catarata%20senil%2C) . Fechade consulta: 21 de febrero de 2025.
- Iqbal, M. Z. & Campbell, A. G. (2022). Adopting Smart glasses responsibly: potencial benefits, ethical, and privacy concerns whit Ray-Ban Stories. *AI and Ethics*, 3(1): 325-327. DOI: 10.1007/s43681-022-00155-7.
- Meta (2021). Ray-Ban and Facebook introduce Ray-Ban Stories, firt- generation Smart glasses. [En línea]. Disponible en: <https://tech.facebook.com/real> (Fang et al., 2026; Muambo et al., 2024)ity-labs/2021/9/ray-ban-and-facebook-introduce-ray-ban-stories-first-generation-smart-glasses/. Fecha de consulta: 9 de febrero de 2025.
- OMS, Organización Mundial de la salud (2023), ceguera y discapacidad visual. [En línea]. Disponible en: <https://www.who.int/es/news-room/fact-sheets/detail/blindness-and-visual-impairment#:~:text=Prevalencia,todav%C3%ADa%20no%20se%20ha%20tratado>. Fecha de consulta: 16 de febrero de 2025.
- Tamayo Zapata, K. J. (2018). Siatema de reconocimiento de billetes para personas con discapacidad visual mediante visión artificial. Tesis de Maestría. Universidad EIA.
- Waisberg, E., ONG, J., Masalkhi, M. Zaman, N., Sarker, P., Lee, A. G. & Tavakkoli, A. (2023). Meta Smart glasses-large language models and the future for assistive glasses for individuals whit vision impairments. *Eye*, 38(6): 1036-1038. DOI: 10.1038/s41433-023-02842-z.

