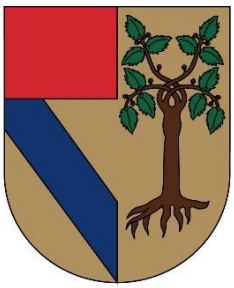


UNIVERSIDAD PANAMERICANA

FACULTAD DE INGENIERÍA

Reconocimiento visual de objetos basado en aprendizaje profundo para asistir a personas con discapacidad visual

TESIS

QUE PRESENTA

Juan Bernardo Calabrese

PARA OBTENER EL GRADO DE

MAESTRÍA EN CIENCIAS

CON RECONOCIMIENTO DE VALIDEZ OFICIAL DE ESTUDIOS DE LA SECRETARIA DE EDUCACION PUBLICA, DE ACUERDO CON EL Nº 2007574 DE FECHA 29 DE JUNIO 2007.

DIRECTOR DE TESIS

Dr. Ramiro Velázquez Guerrero

AGUASCALIENTES, AGS, NOVIEMBRE 2020.

Índice

Resumen

Enmarcada en el campo de las tecnologías de asistencia (TA), esta tesis presenta un novedoso sistema electrónico cuyo objetivo es proporcionar reconocimiento visual de objetos comunes con el fin de facilitar su búsqueda a las personas con discapacidad visual (DV). El sistema consta de dos componentes principales: una cámara de video en miniatura y un sistema embebido (SoM). El primero se usa en los lentes del usuario y adquiere video en tiempo real del espacio cercano, mientras que el segundo se coloca en el cinturón y ejecuta métodos basados en aprendizaje profundo y algoritmos espaciales que procesan el video proveniente de la cámara realizando la detección y el reconocimiento de objetos así como su posicionamiento en el espacio circundante. El dispositivo desarrollado proporciona frases descriptivas audibles como retroalimentación al usuario que involucran los objetos reconocidos y su posición referenciada a su mirada. Los resultados experimentales obtenidos con el prototipo desarrollado han demostrado una identificación precisa y confiable de objetos en tiempo real con una tasa de reconocimiento del 86% y un tiempo de cómputo promedio de 215 ms. El sistema propuesto es capaz de reconocer los 91 objetos ofrecidos por la base de datos COCO más cuatro objetos personalizados. Además, esta tesis introduce una metodología simple y escalable para la utilización de bases de datos de imágenes y entrenamiento de redes neuronales convolucionales (CNN) para agregar objetos al sistema y aumentar su repertorio. También se demuestra que los entrenamientos más exhaustivos que involucran 400 imágenes logran tasas de reconocimiento del 89%, mientras que entrenamientos rápidos con solo 40 imágenes logran tasas de reconocimiento aceptables del 55%.

1. Introducción

A nivel mundial, la Organización Mundial de la Salud (OMS) estima que hay alrededor de 235 millones de personas con discapacidad visual severa o ceguera completa, cuya patología no se puede corregir con el uso de lentes estándares o cirugía [1]; las mismas consideraciones y cifras similares también se han reportado en [2].

La discapacidad visual (DV) interfiere con la capacidad de la persona para realizar actividades cotidianas como comprensión del entorno, movilidad urbana, lectura, acceso a computadoras, búsqueda de objetos, entre otras [3,4]. Tratar de adaptarse al mundo, en mayor o menor medida, es un desafío constante. Numerosos trabajos han abordado la comprensión del medio ambiente [5,6] y el problema de la navegación urbana [7,8] mediante el desarrollo de dispositivos de asistencia basados en sensores inteligentes o visión artificial [9]; otros han propuesto soluciones para la lectura [10,11] y el acceso informático [12,13]. Sin embargo, pocos trabajos se han centrado en asistir a las personas con DV con la tarea de encontrar objetos de uso cotidiano. La detección y el reconocimiento de objetos en una escena podría facilitar la vida de las personas con DV: encontrar y alcanzar elementos en el espacio circundante aumenta la calidad de vida y la seguridad. No saber qué hay alrededor puede generar frustración y ansiedad resultando en situaciones peligrosas que podrían provocar tropiezos, caídas, quemaduras y lesiones. La detección e identificación automática de objetos para las personas con DV requiere un enfoque flexible, adaptable y computacionalmente eficiente que aprenda continuamente y aumente gradualmente su conocimiento.

Los avances recientes en redes neuronales y el aprendizaje profundo han contribuido a los avances en el campo de la visión por computadora. Las redes neuronales profundas (DNN), especialmente las redes neuronales convolucionales (CNN), han demostrado ser muy eficaces en áreas como el reconocimiento y la clasificación de imágenes [14, 15]. En particular, las arquitecturas VGG16 (CNN de 16 capas) y VGG19 (CNN de 19 capas) se han utilizado ampliamente para estas tareas, ya que requieren una cantidad moderada de tiempo de entrenamiento; sin embargo, no son eficientes para aplicaciones en tiempo real con requerimientos de altas velocidades de procesamiento.

Para palear la carga del procesamiento computacional que limita la velocidad de clasificación de las CNN, la R-CNN (CNN basada en regiones) fue la primera opción propuesta. Ésta selecciona varias regiones de la imagen y luego hace uso de las CNN para extraer características de cada región [16]. Sin embargo, la selección de varias regiones requiere que la CNN realice un número significativo de cálculos; en consecuencia, la carga computacional hace que las R-CNN sean ineficientes para aplicaciones en tiempo real.

Fast R-CNN mejora la operación de R-CNN al calcular la imagen como un todo. En lugar de enviar el conjunto de regiones a la CNN, Fast R-CNN proporciona la imagen de entrada directamente a la CNN para generar un único mapa de características convolucionales. A partir de este mapa, las regiones propuestas se identifican utilizando algoritmos de búsqueda selectiva. Fast R-CNN reduce drásticamente los tiempos de entrenamiento y detección en comparación con R-CNN [17]. Aun así, su principal inconveniente es que requiere la generación de muchas regiones propuestas para obtener un reconocimiento de objetos preciso. Por tanto, el cuello de botella de esta arquitectura es el algoritmo de búsqueda selectiva. Su sucesor, Faster R-CNNs reemplaza la búsqueda selectiva con una red de regiones propuestas [18]. Este método permite reducir los tiempos de procesamiento sin perder precisión en la detección. La Figura 1 compara el tiempo de procesamiento de estas tres arquitecturas basadas en R-CNN para realizar el reconocimiento de objetos en una imagen. Nótese que Faster R-CNN supera a sus predecesores; de hecho, mientras que R-CNN y Fast R-CNN tardan 50 y 2.3 s (este último valor corresponde al 4.6% del tiempo empleado por las R-CNN), respectivamente, Faster R-CNN ejecuta la misma tarea en solo 0.2 s, lo que representa solo el 0.4% del tiempo consumido por las R-CNN [17].

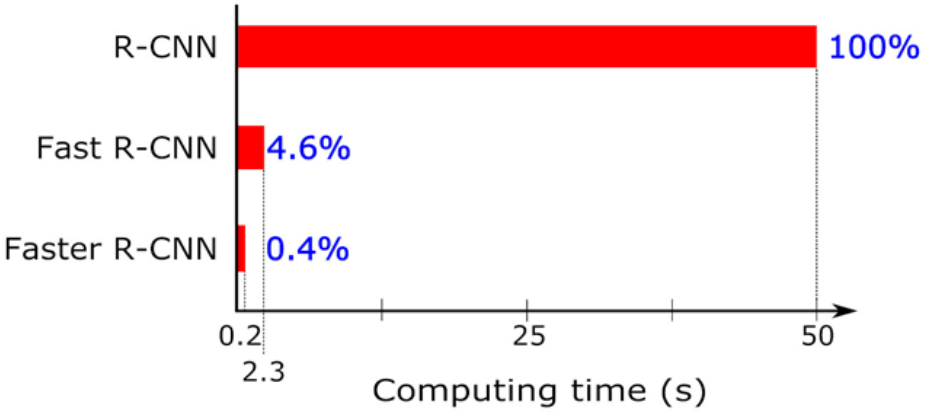


Figura 1. Comparación del rendimiento de las arquitecturas R-CNN en tareas de reconocimiento de objetos.

Esta tesis presenta un enfoque funcional para el diseño e implementación de un sistema basado en aprendizaje profundo simple, de bajo costo y eficiente, enfocado en la búsqueda de objetos para personas con DV, mediante el empleo de Faster R-CNN. Además del reconocimiento de objetos, el sistema realizado ayuda al usuario a posicionar los objetos en el espacio circundante integrando un script en el código principal. En las CNN, mientras más imágenes se utilicen para el entrenamiento, se obtendrá un mejor resultado; sin embargo, en este trabajo, entrenamientos rápidos que brindan buenos resultados (tasas de reconocimiento aceptable del 55%) son posibles con solo 40 imágenes.

La novedad de esta propuesta radica en presentar una solución completa que integra estructuras de aprendizaje profundo en hardware portátil que es de bajo costo, confiable, de alto rendimiento y verdaderamente asequible para los usuarios finales. Además, la plataforma exhibe una gran flexibilidad al permitir la adición de nuevos objetos a su base de datos interna de manera sencilla. Las capacidades de reconocimiento del dispositivo se pueden aumentar progresivamente y responder a las necesidades cambiantes del usuario. A nuestro conocimiento, un sistema de este tipo dedicado a abordar las necesidades de la comunidad con DV no ha sido previamente reportado en la literatura.

Esta tesis se estructura de la siguiente manera. En la sección 2 proporciona una revisión detallada de la literatura sobre sistemas que brindan detección y reconocimiento de objetos, en particular, aquellos dedicados a ayudar a las personas con DV. En la sección 3 se describen los componentes del sistema y detalla los

principales conceptos involucrados. En la sección 4 se presentan los resultados experimentales obtenidos con el sistema propuesto mientras que en la sección 5 se presenta una discusión de los resultados experimentales obtenidos y se comparan con los obtenidos en otros trabajos científicos. Finalmente, en la sección 6 concluye la tesis resumiendo las principales contribuciones y proporcionando las perspectivas de trabajo futuro.

2. Trabajos relacionados

En este capítulo se presenta una descripción general de dispositivos de TA destinados a asistir a personas con DV proporcionando, en un principio, la clasificación de las diferentes tecnologías empleadas en estas aplicaciones. Posteriormente, se analizan varios métodos de reconocimiento visual, clasificándolos en función de su aplicación.

2.1. Estado del arte en dispositivos innovadores de TA para ayudar a las personas con DV.

Recientemente, se han reportado en la literatura científica diferentes dispositivos portátiles para ayudar a personas con DV, cuyo objetivo es mejorar su calidad de vida y reducir la probabilidad de accidentes. Existen diferentes tecnologías disponibles para el reconocimiento de objetos, la detección de obstáculos, la asistencia a la navegación, etc. En particular, la clasificación de las principales tecnologías empleadas para esta tipología de dispositivos, en función de sus aplicaciones, se resume en la Tabla 1.

Tabla 1. Resumen de las principales tecnologías utilizadas en dispositivos TA

Tecnología	Aplicación
Sistema de reconocimiento visual.	Detección de objetos, Detección de rostros, Navegación, Posicionamiento, Control de acceso, Reconocimiento de texto.
Tecnología RFID	Detección de objetos, Posicionamiento, Control de acceso.
Sistema de navegación GPS.	Navegación, posicionamiento.

Sistema de detección de ultrasonidos.	Navegación, Control de acceso, Detección de obstáculos.
LIDAR / Sensores de distancia ópticos	Navegación, Detección de obstáculos.
Interfaces vibrotáctiles.	Interfaz de usuario.
Interfaces basadas en audio.	Interfaz de usuario.

A continuación, se presenta un resumen de trabajos científicos que involucran las tecnologías reportadas en la tabla anterior; en varios casos, estos sistemas integran múltiples tecnologías para incrementar sus funcionalidades o mejorar su fiabilidad.

V. Meshram et al., en [19], propuso un dispositivo innovador de TA, llamado "Nav-Cane", para ayudar a las personas con DV en la orientación y la navegación en ambientes interiores o exteriores, identificando objetos u obstáculos presentes en el camino del usuario. El dispositivo desarrollado está equipado con un lector de identificación por radiofrecuencia (RFID), sensores ultrasónicos colocados a diferentes alturas, un receptor GPS, un sensor de inercia y un sensor de agua, todos integrados dentro de una estructura en forma de bastón. El lector RFID se utiliza para reconocer objetos previamente etiquetados, así como los sensores ultrasónicos brindan retroalimentación directa al usuario, a través de avisos táctiles producidos por un motor de vibración, relacionados con la presencia de obstáculos a diferentes niveles (por ejemplo, pie, rodilla, cadera, pecho). Además, el sensor de agua advierte al usuario de la presencia de pisos mojados y se agrega un botón de alarma para alertar, a través de un correo electrónico o Servicio de Mensajes Cortos (SMS), a los equipos de rescate indicando las coordenadas GPS del usuario. Los resultados experimentales llevados a cabo en 80 personas con DV en ambiente controlados, demostraron la efectividad del Nav-Cane en la detección de obstáculos, para bajar o subir escaleras, así como para ayudar al usuario a identificar objetos.

De manera similar, en [20], los autores propusieron "The Assistor", un bastón inteligente para ayudar a las personas invidentes o con DV a reconocer obstáculos y así navegar hasta su destino. Este dispositivo emplea tres sensores ultrasónicos, dos en la parte superior y dos en la parte inferior para cubrir un ángulo más amplio y una cámara miniaturizada integrando un potente procesador que admite varios

protocolos de comunicación para transmitir imágenes adquiridas a un microcontrolador.

Los datos de los sensores se envían a través de comunicación Bluetooth a un teléfono inteligente, donde se procesan para activar el servomotor utilizado para mover el bastón. El receptor GPS del teléfono inteligente y la aplicación de mapas de Google guían al usuario hacia su destino, donde el sensor identifica obstáculos utilizando el algoritmo de Speeded Up Robust Features (SURF) , soportado por una base de datos imagen-objeto.

En [21], los autores presentaron un sistema de detección de obstáculos basado en múltiples dispositivos de sonar, constituidos por varios sensores de ultrasonido y distribuidos para detectar una gran área del campo de visión (FOV) del usuario; un actuador está asociado a cada sensor, representado por un motor vibratorio e instalado para proporcionar retroalimentación vibrotáctil al usuario con respecto a la posición del obstáculo. El dispositivo es gestionado por un microcontrolador PIC18F6720 que recoge los datos de los sensores ultrasónicos y los procesa para accionar los actuadores, según una calibración previa realizada en las condiciones del usuario. Cinco usuarios diferentes probaron el dispositivo desarrollado. Los resultados experimentales indican una reducción del 50% del tiempo requerido por el usuario para atravesar una serie de obstáculos.

S. Chen *et al*. propusieron un sistema portátil inteligente para el reconocimiento de imágenes que adopta la nube y el procesamiento cooperativo local [22]. Específicamente, un servidor en la nube realiza tareas de procesamiento de imágenes mientras que la unidad de procesamiento local carga las imágenes en dicho servidor y recibe los resultados del procesamiento. Por tanto, se pueden emplear procesadores de bajo costo y recursos limitados para desarrollar el dispositivo portátil, reduciendo así su costo total. El dispositivo portátil propuesto incluye una micro-cámara, un sensor ultrasónico y un sensor infrarrojo instalados en el armazón de los lentes; se utiliza un Raspberry Pi como procesador local, proporcionando al dispositivo conectividad inalámbrica (WiFi o 4G) para compartir las imágenes con el servidor en la nube aprovechando su potencia de cómputo paralelo y capacidad de almacenamiento. Un algoritmo combina las imágenes capturadas con los datos proporcionados por los sensores ultrasónicos e infrarrojos para extraer el "punto de interés" útil para el reconocimiento. Los resultados

experimentales demostraron la efectividad del dispositivo propuesto en la detección de rostros, texto y objetos, en escenarios reales. En [23], los autores propusieron el sistema portátil "Obstacle Stereo Feedback" (OSF) para ayudar a la navegación de usuarios con DV; el sistema implementa un algoritmo consensuado de muestreo aleatorio (RANSAC) para elaborar la nube de puntos adquirida y detectar obstáculos colocados a lo largo del camino del usuario. Además, las funciones de transferencia relacionadas con cabeza (HRTF) se emplean para proporcionar una representación acústica de los obstáculos en función de sus coordenadas 3D.

Neto *et al.* en [24] presentaron un sistema de reconocimiento facial para ayudar a usuarios con DV; el sistema desarrollado emplea una barra de sensores basada en un Kinect instalada en un casco y un algoritmo KNN, basado en descriptores orientados a histogramas comprimidos por el método del componente principal. Los resultados experimentales demuestran que el algoritmo de detección propuesto requiere menores recursos computacionales en comparación con otras técnicas reportadas en la literatura, manteniendo una excelente precisión aún bajo diferentes condiciones operativas, como fondo, iluminación y punto de vista.

Katzschmann *et al.,* en [25], presentaron el "Array of Lidarsand Vibrotactile Unit" (ALVU), un dispositivo portátil, pensado para usuarios invidentes o con DV, que detecta obstáculos y sus límites físicos. El dispositivo incluye un cinturón y una correa háptica; el primero está equipado con sensores de distancia de tiempo de vuelo, para medir la distancia entre el usuario y los obstáculos; mientras que el segundo proporciona retroalimentación al usuario mediante una serie de motores vibratorios colocados en el abdomen del usuario.

Se pueden encontrar en la literatura otras aplicaciones relevantes que involucran el reconocimiento de objetos en tiempo real: Chen *et al.* introdujeron en [26] el "Glimpse system": un sistema de reconocimiento de objetos en tiempo real para teléfonos inteligentes, que se ejecuta en servidores externos. Los experimentos con señales de tráfico muestran una precisión de reconocimiento del 75 al 80%.

Viola y Jones propusieron en [27] un marco de detección de rostros capaz de procesar imágenes de manera extremadamente rápida y lograr altas tasas de detección (95%). Jauregui y colaboradores abordaron la identificación de puertas

para la navegación de robots en espacios interiores [28]. Al emplear un algoritmo de tres etapas, lograron tasas de reconocimiento del 98%.

Pocos sistemas se han centrado en las necesidades de las personas con DV. Niu et al. describen en [29] un sistema portátil que detecta picaportes y manos humanas para ayudar a las personas invidentes a localizar y utilizar puertas. Panchal et al. presentaron en [30] un nuevo enfoque para reconocer texto a partir de imágenes de escena y convertirlo en habla para que pueda ayudar a las personas con DV. Jabnoun y colegas reportaron en [31] un sistema de detección de objetos basado en los algoritmos SIFT (Scale Invariant Features Transform) y SURF (Speeded Up Robust Features) para asistir a personas con DV en la navegación. Ciobanu et al. introdujeron en [32] un método para detectar escaleras en interiores mediante el empleo de sensores IMU (inertial measurement unit) y el procesamiento de imágenes de profundidad con el fin de ayudar a las personas con DV en entornos desconocidos.

En este contexto, este trabajo da un paso adelante en el campo del reconocimiento de objetos para asistir a personas con DV en la detección de objetos de uso diario como picaportes, enchufes, bastones, interruptores de luz, entre muchos otros. El sistema propuesto captura video en tiempo real, localiza los objetos de interés, los reconoce, rastreando fotograma a fotograma y, finalmente, proporciona retroalimentación descriptiva audible al usuario.

2.2. Resumen de aplicaciones y métodos para el reconocimiento visual

El reconocimiento de objetos se ha convertido en un tema importante en el campo de la visión por computadora. Se ha explorado exitosamente en video vigilancia [33], navegación robótica [34], imágenes médicas [35], hogares inteligentes [36] e incluso en turismo [37]. En esta sección, se informan y detallan algunos métodos utilizados para el reconocimiento de objetos, proporcionando también un análisis comparativo entre los algoritmos reportados.

En [33], los autores propusieron un método innovador para determinar el rendimiento de los algoritmos de reconocimiento de objetos en video, destacando características específicas del método en particular, como la división o fusión de regiones. El método se basa en la comparación entre la salida del algoritmo de

reconocimiento y el segmento dividido correcto extraído con una frecuencia de muestreo de 1 fotograma /s. Similarmente, Lu *et al.*, en [38], introdujeron un marco de detección de objetos en tiempo real para video, empleando la red You Only Look Once (YOLO), con un método de convolución mejorado para acelerar la elaboración y, por lo tanto, la detección de objetos. Mediante un pre-procesamiento, se eliminan los efectos del fondo, así como el ruido. Los resultados experimentales indican que el método propuesto obtiene mejores rendimientos en comparación con el algoritmo YOLO inicial, alcanzando mayor velocidad de detección y precisión.

Además, los algoritmos de reconocimiento de objetos pueden encontrar aplicaciones para sistemas de navegación en vehículos autónomos (AV) y campos robóticos; Hernández *et al.* propusieron un sistema de reconocimiento de objetos, basado en el método de clasificación Support Vector Machine (SVM) aplicado en imágenes RGB [34]. Se han probado dos enfoques de segmentación basados en descriptores de formas geométricas y el método de la bolsa de palabras, respectivamente.

Estos algoritmos se pueden utilizar también para aplicaciones de imágenes médicas, ayudando a usuarios con DV a realizar tareas que de otro modo no podrían realizar; por ejemplo, en [35], los autores propusieron una aplicación móvil para permitir a usuarios invidentes leer un texto. El proyecto "Camera Reading for Blind People" utiliza Reconocimiento Óptico de Caracteres (OCR) y síntesis texto-discurso (TTS), integrado en un teléfono inteligente, para adquirir imágenes de un texto y sintetizar vocalmente el texto reconocido.

Además, los sistemas automatizados para hogares inteligentes pueden explotar ampliamente los sistemas de reconocimiento de objetos; Baeg et al. desarrollaron un entorno doméstico inteligente para robots de servicio, equipado con una cámara, un lector RFID y un módulo de comunicación inalámbrica [36].

El sistema de reconocimiento visual también se puede utilizar para reconocer lugares, monumentos, estatuas, pinturas, etc., con el fin de proporcionar información y datos a los turistas. En [37], los autores propusieron un sistema de visión móvil para el reconocimiento automático de objetos aplicado a las imágenes adquiridas mediante un teléfono con cámara. El sistema desarrollado permite determinar lugares de interés turístico, para brindar información detallada

relacionada con la arquitectura, la historia o el contexto cultural de relevancia histórica o artística.

Trabelsi *et al.*, en [39], desarrollaron un novedoso algoritmo multimodal para ayudar a usuarios con DV en el reconocimiento de objetos en un ambiente interior, empleando imágenes RGB-D (Red Green Blue-Depth) con una nueva representación de valor complejo, con el fin de superar las limitaciones de las técnicas tradicionales. Se han propuesto dos métodos para categorizar objetos; Multi-Label - Complex-Valued Neural Networks (ML-CVNN), basado en un método de agrupamiento adaptativo, que se aplica a la resolución de problemas de múltiples etiquetas. Este último, llamado L-CNN, usa un CNN para cada etiqueta considerada para obtener un vector multi-etiqueta. Los resultados experimentales demuestran la eficiencia de las técnicas propuestas basadas en imágenes RGB-D en comparación con métodos existentes, como RGB ML-Real-Valued Neural Network (RVNN), Depth ML-RVNN y RGBD ML-RVNN, en la clasificación de objetos.

Del mismo modo, Malūkas et al. introdujeron un sistema de navegación en tiempo real para personas invidentes o con DV, empleando un marco de segmentación basado en un algoritmo de red neuronal convolucional (CNN) profunda para reconocer objetos y características en una imagen [40]. Se han probado tres algoritmos CNN diferentes (ie AlexNet, GoogLeNet y VGG-Visual Geometry Group Net), obteniendo los mejores rendimientos en la segmentación con la red neuronal VGG16 (16 capas) y alcanzando una precisión de 96.1 ± 2.6% en el reconocimiento de caminos, estructuras y límites de caminos.

Jayakanth propuso un algoritmo en tiempo real para el reconocimiento de objetos en ambientes interiores, como puertas, escaleras y letreros [41]; el algoritmo se basa en un enfoque de aprendizaje por transferencia para implementar un modelo de aprendizaje profundo entrenado con AlexNet. Además, en el marco propuesto se han probado diferentes técnicas de extracción de características de textura (patrón binario local-LBP, características de imagen estadística binarizada-BSIF y cuantificación de fase local-LPQ), seguidas de un clasificador de aprendizaje automático (es decir, K-vecinos más cercanos-KNN, Naive Bayes-NB y SVM) para la identificación y clasificación de objetos. Los resultados de la prueba demuestran que los extractores de textura BSIF y LPQ funcionan de manera excelente en el reconocimiento de objetos; específicamente, el primero en conjunto con el

clasificador KNN obtiene una precisión del 98.4%, así como el segundo produce el 98.9% y el 98.4% cuando opera con los clasificadores SVM y KNN, respectivamente.

Además, Jabnoun et al., describieron en [42], una herramienta visual para personas con DV, basada en un algoritmo de reconocimiento de objetos que permite determinar las diferencias entre cuadros de video; específicamente, el algoritmo emplea el método "Real-Valued Local Dissimilarity Map" (RVLDM) como medida de la diferencia de los cuadros, y el extractor "Scale-Invariant Features Transform" (SIFTS), para determinar los objetos representados en diferentes marcos. Al comparar el método propuesto con enfoques de sustitución visual similares, se han demostrado rendimientos óptimos, en términos de velocidad computacional en diferentes condiciones operativas, como diferentes puntos de vista, presencia de oclusión, rotación del marco y diferente iluminación.

Además, en [43], los autores describieron un novedoso método de reconstrucción de objetos 3D, basado en una red neuronal artificial híbrida modificada, obteniendo un relleno más preciso de imágenes de objetos parciales y reduciendo el ruido del proceso en comparación con el algoritmo YOLOv3. Además, la reconstrucción obtenida es más estable que los resultados obtenidos con otras técnicas de reconstrucción.

En la Tabla 2, se resumen los rendimientos de las implementaciones de algoritmos de reconocimiento de objetos comunes previamente mencionadas. Como es evidente, los entornos YOLO alcanzan una alta velocidad de procesamiento pero imponen fuertes restricciones espaciales en las predicciones de los cuadros delimitadores, lo que limita la capacidad del algoritmo para discernir objetos cercanos [44].

Además, los algoritmos de CNN basados en regiones, y específicamente los de R-CNN más rápidos, representan una excelente compensación entre la velocidad de operación, la precisión y la utilización de recursos [45].

Tabla 2. Resumen del desempeño de las implementaciones de algoritmos de reconocimiento de objetos comunes.

Algoritmo	Arquitectura	Precisión media [%]	Media de cálculo [ms]	Base de datos
A.C. Hernández [34]	SVM	78.34	n.a.	NYU Depth Dataset V2 [46]

S. Lu [38]	Fast YOLO	88.45	22	Personalizado
Trabelsi [39]	ML-CVNN	87.2	n.a	RGBD [47]
U. Malūkas [40]	FCN-VGG16	96.1	105	ImageNet [48]
K. Jayakanth [41]	CNN+SVM	100	451	MCindoor20000 [49]
J. Redmon [50]	YOLOv3	57.9	51	COCO
S. Ren [45]	Faster R-CNN	73.2	198	COCO

2. Descripción del sistema

Para abordar el problema objeto de esta tesis, hemos desarrollado un dispositivo de tecnología de asistencia (TA) que se puede portar en la vestimenta (es decir, es wearable) con capacidad para reconocer objetos de uso cotidiano y que permite al usuario conocer su presencia y ubicación en el espacio circundante.

La figura 2 muestra la representación conceptual del dispositivo TA y su principio de funcionamiento. El sistema comprende dos elementos principales: una cámara en miniatura y un sistema en módulos (SoM); las imágenes adquiridas por la cámara se procesan en tiempo real por el SoM para detectar objetos de uso diario. Cuando se detecta un objeto, el SoM emite el nombre del objeto a través de un altavoz integrado que permite al usuario saber su presencia y ubicación en el espacio cercano.

Las siguientes subsecciones detallan los elementos del sistema.

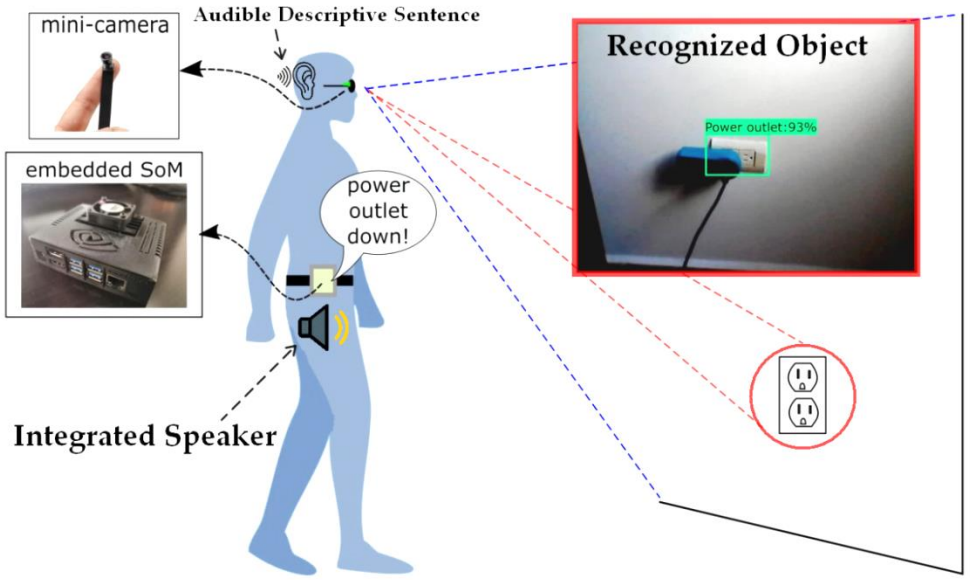


Figura 2. Dispositivo TA portable para el reconocimiento de objetos. En este ejemplo, la cámara captura una imagen de una pared que contiene una toma corriente, los algoritmos que se ejecutan en el SoM lo detectan e informan al usuario sobre su presencia y ubicación a través de audio.

3.1. Hardware

La cámara miniatura empleada es un sensor CMOS RGB (modelo 6986154272705, fabricado por Hamswan Company, Shenzhen, China) con un precio indicativo de 30 - 35 USD (Figura 3a). Ésta proporciona imágenes con una resolución de 1280 x 960 píxeles y un campo de visión (FOV) de 120 grados (como referencia, el FOV humano es de 190 grados [51]). Sus dimensiones (12.5 mm x 12.5 mm x 17 mm) y masa (12 g) permiten colocarla fácilmente en el armazón de los lentes, de tal forma que los usuarios pueden llevarla totalmente integrada a los lentes, sin siquiera notarlo. La cámara en miniatura se interconecta directamente con un módulo de comunicación y procesamiento local, que adquiere las imágenes y las transmite de forma inalámbrica a la unidad de procesamiento principal (SoM), mediante comunicación WiFi (Figura 3a); de esta forma, no se requieren cables, permitiendo al usuario una máxima libertad de movimiento. Su tamaño compacto, bajo costo y alimentación USB la hacen ideal para esta aplicación.

El SoM utilizado es el sistema embebido Jetson Nano (fabricado por NVIDIA Co., Santa Clara, EE. UU.) Con un precio indicativo de 120 - 130 USD. Su RAM LPDDR4 (Low-Power Double Data Rate) de 4 Gb, la unidad de procesamiento de gráficos (GPU) basada en la arquitectura NVIDIA Maxwell de 128 núcleos y la unidad central de procesamiento (CPU) ARM A57 de cuatro núcleos que se ejecuta hasta 1.4 GHz, permiten la ejecución de algoritmos basados en inteligencia artificial (IA) en tiempo real [52,53]. La placa del SoM incluye varias interfaces, como DisplayPort, HDMI, cuatro puertos USB-Universal Serial Bus 3.0, dos conectores CSI-Camera Serial Interface, Gigabit Ethernet, una ranura para tarjeta Wifi M.2 y un conjunto de GPIO colocados en la lateral del tablero, lo que lo hace ideal para una variedad de aplicaciones de IA en vehículos autónomos y robótica (Figura 3b). Sus dimensiones (70 mm x 45 mm x 25 mm) y masa (240 g) lo hacen portátil; por ejemplo, los usuarios pueden sujetarlo al cinturón, como se muestra en la Figura 2. El SoM se

alimenta con 5 V CC provistos por un conector micro-USB. El consumo de energía del SoM está en el rango de 5 - 10 W, según la carga computacional requerida del módulo, y este consumo está garantizado por un paquete de baterías NiMH de 4600 mAh recargadas continuamente por medio de un sistema de recolección de energía solar. Pruebas experimentales revelan que se puede obtener una autonomía energética de 1.6 h hasta 3.2 h en función del modo de operación del SoM, el cual se puede potenciar aún más, en función del nivel de iluminación solar.

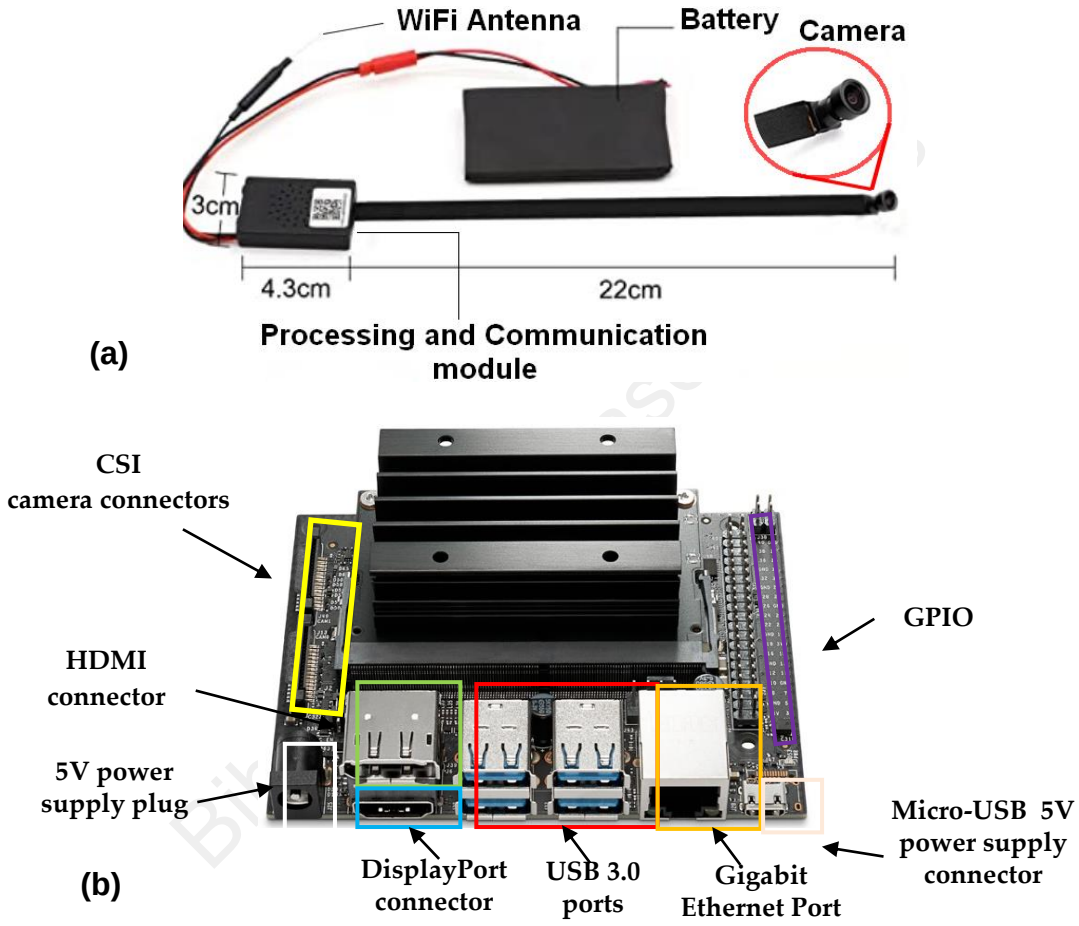


Figura 3. La cámara utilizada para el sistema TA propuesto con las secciones principales resaltadas (a); vista superior del módulo electrónico Jetson Nano con sus interfaces principales indicadas (b).

La cámara CMOS RGB empleada se conecta a través del puerto USB al SoM, que lo reconoce automáticamente sin necesidad de instalar software adicional (es decir, es modo plug and play). Las cámaras y el hardware compatibles con el Jetson Nano se pueden encontrar en [54, 55].

Finalmente, se integró un altavoz estándar al SoM usando sus puertos de entrada y salida. Se prefirió un altavoz a los auriculares para evitar la obstrucción de la audición ambiental; de hecho, las personas con DV dependen en gran medida de las señales ambientales para navegar y orientarse en el espacio [3].

El módulo Jetson Nano integra varias soluciones para optimizar su consumo de energía, adaptándose a la aplicación específica. En particular, tiene dos modos de funcionamiento con diferente consumo de energía, el modo 0 (también llamado modo MaxN) y el modo 1 (también llamado modo 5 W). En el primero, se habilita el consumo de energía de 10 W para obtener el máximo rendimiento, el segundo, el consumo de energía se limita a solo 5 W, al restringir las frecuencias de reloj de la memoria, CPU, Unidad de Procesamiento Gráfico (GPU) y el número de núcleos activos [56].

Específicamente, el modo 1 tiene frecuencias máximas de reloj de CPU y GPU limitadas a 918 MHz y 640 MHz, respectivamente y solo dos núcleos activos. El modo de funcionamiento se puede cambiar mediante el comando nvpmode, pasando como parámetro el identificador de la modalidad seleccionada. Habilitando el modo 1, el tiempo de reconocimiento de objetos aumentó en comparación con la elaboración en modo MaxN, obteniendo un tiempo de procesamiento medio de 360 ms, pero dejando inalterada la precisión media (es decir, 86 %).

El sistema de reconocimiento desarrollado está equipado con un acelerómetro MEMS tri-axial (modelo MMA8451Q, fabricado por NXP, Eindhoven, Netherland) que se utiliza para detectar la velocidad del usuario y, por ende, adaptar dinámicamente el consumo de energía del SoM según la condición del usuario. Específicamente, si el usuario está parado o se mueve lentamente, no se necesita una alta velocidad de reconocimiento, por lo que el modo 1 se habilita; por el contrario, si el usuario camina rápido, se requiere una detección rápida de objetos, por lo que en esta condición se habilita el modo 0. La lógica de administración de energía detecta continuamente la velocidad del usuario y, si esta última es superior a 2 m/s durante un intervalo de tiempo superior a 5 s, se establece el modo 0; de lo contrario, el sistema se configura en modo 1. De esta forma, se obtiene una reducción en el consumo energético del sistema desarrollado, en comparación con el funcionamiento continuo del dispositivo en modo MaxN, dejando, desde un punto de vista práctico, inalterada la funcionalidad.

Por tanto, la autonomía energética del sistema portátil TA desarrollado se incrementa, dentro del rango entre valores extremos en los que el sistema consume constantemente 5 W o 10 W. La placa breakout MMA8451Q se colocó dentro de la tapa que contiene la placa Jetson Nano, fijada en el cinturón del usuario (como se muestra en la Figura 2), ésta es la posición ideal para detectar los movimientos del cuerpo [57].

Con el fin de verificar los posibles desarrollos comerciales existentes similares a nuestra propuesta, se llevó a cabo una investigación exhaustiva de la literatura y en la web. Recientemente se ha presentado en el mercado un dispositivo portátil comercial para personas con DV llamado Horus [58]. Éste se basa en la plataforma informática de inteligencia artificial integrada NVIDIA Jetson y se compone de un auricular portátil con cámaras y una unidad de bolsillo que contiene un procesador y una batería de larga duración.

Horus se presenta en [58] como un dispositivo portátil que observa, comprende y describe el entorno a la persona que lo usa, proporcionando información útil; es capaz de leer textos, reconocer rostros, objetos y el usuario puede activar cada funcionalidad a través de un conjunto de botones ubicados tanto en el auricular como en la unidad de bolsillo.

La arquitectura, los algoritmos y la implementación de software del dispositivo propuesto en esta tesis están optimizados y son más simples que el prototipo Horus. Una comparación final de costos muestra definitivamente que la solución propuesta en esta tesis podría ser más competitiva considerando también un desarrollo comercial; de hecho, el costo total del prototipo propuesto se ha evaluado en solo 200 USD, mientras que el precio comercial propuesto para el dispositivo Horus es de 2,000 USD, diez veces más costoso.

3.1.1. Diseño del sistema de recolección de energía solar para extender la autonomía del dispositivo TA

Adicionalmente, se ha desarrollado un sistema de recolección de energía solar portátil para suministrar energía al dispositivo TA, lo que le permite extender su autonomía energética. En particular, dos paneles solares mono-cristalinos flexibles de 6W (modelo HX160-220P, fabricado por Huaxu Energy Co., Shenzen, China),

conectados en paralelo, se han colocado en la parte posterior de una camisa de tela mediante botones de clip metálicos para facilitar su instalación y retiro.

Específicamente, los paneles son células solares laminadas de tereftalato de polietileno (PET), caracterizadas por un voltaje de circuito abierto de 7.2 V, corriente de cortocircuito de 1100 mA, voltaje pico de 6 V, corriente pico de 1000 mA, eficiencia de conversión del 19.5% y dimensiones: 160 mm x 220 mm x 2.8 mm (Figura 4). Las células solares están interconectadas con un convertidor de voltaje buck-boost S18V20F6 (fabricado por Pololu Co., Las Vegas, EE. UU.), que presenta un amplio rango de entrada (de 2.9 a 32 VCC), un voltaje de salida fijo de 6 VCC con una precisión del 4%, una corriente de salida máxima de 2 A y eficiencia típica entre 80% y 90%. Además, la placa está equipada con protección de voltaje inverso (hasta 30 V), protección contra sobrecorriente y circuito de apagado por sobrecalentamiento. La carga extraída por los paneles solares se almacena en el paquete de baterías NiMH de 4600 mAh, 6 VCC (fabricado por Vapextech UK Ltd, Kent, Reino Unido), que se utiliza para suministrar energía al dispositivo de TA (como se muestra en la Figura 4). Dado que el paquete de baterías de NiMH cuando está completamente cargado alcanza los 6 VCC, el valor de voltaje es incompatible con el rango de suministro de energía de la placa Jetson Nano, un convertidor reductor, basada en el controlador síncrono MP2315 (fabricado por Monolithic Power Company, Kirkland, WA, EE UU.), se ha empleado para proporcionar el voltaje estabilizado de 5 VCC requerido por el SoM (Figura 4). El circuito integrado MP2315 se caracteriza por un amplio rango de voltaje de entrada (de 4.5V a 24 V), una corriente de carga máxima de 3A y alta eficiencia (97%).

La batería y el módulo electrónico se colocaron en bolsillos realizados en la parte interna de la prenda, mientras que el cableado se cosió sobre la tela.

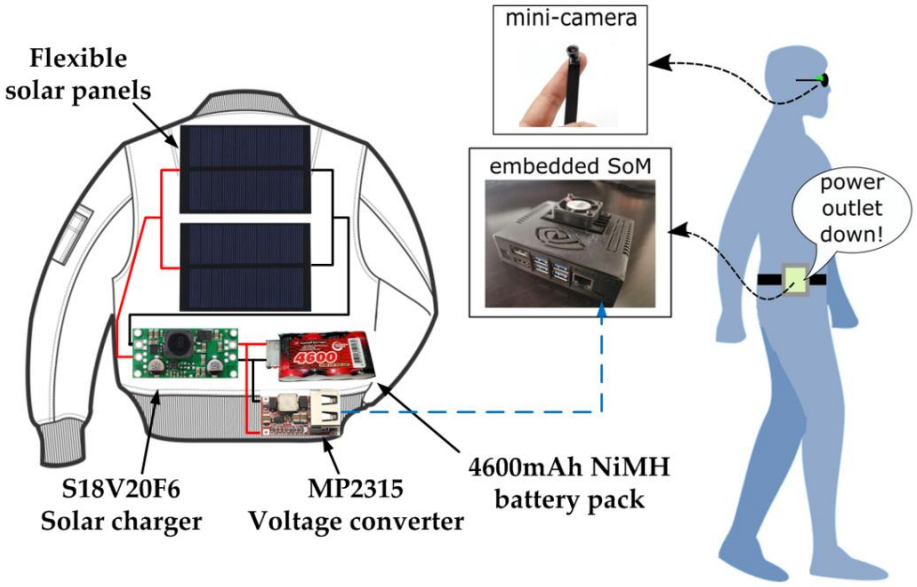


Figura 4. Diseño del sistema de recolección solar portátil integrado con el dispositivo TA desarrollado.

El sistema de captación solar debe ser usado por el usuario algunas horas antes de conectar el dispositivo TA para cargar completamente el paquete de baterías NiMH. Posteriormente, éste último se utiliza para alimentar el dispositivo de reconocimiento de objetos, asegurando así una autonomía energética, en ausencia total de fuentes luminosas, comprendida entre los dos valores límite indicados en las ecuaciones (1) y (2), relacionados con los casos en los que el SoM Jetson Nano está continuamente configurado en modo 0 (modo MaxN, corriente absorbida 2A) y modo 1 (modo 5W, corriente absorbida 1 A), respectivamente.

$$Energy\ Autonomy_{\ mode\ 0} = \frac{Battery\ Capacity\ [mAh] \times (1 - Discharge\_Margin)}{Absorbed\ current\ [mA]} = \frac{4600mAh \times 0.7}{2000mA} = 1.6\ h, \quad (1)$$

$$Energy\ Autonomy_{\ mode\ 1} = \frac{Battery\ Capacity\ [mAh] \times (1 - Discharge\_Margin)}{Absorbed\ current\ [mA]} = \frac{4600mAh \times 0.7}{1000mA} = 3.2h, \quad (2)$$

Donde *Discharge_Margin* es el límite porcentual para la descarga del paquete de baterías NiMH de 4600 mAh (típicamente 30%). Sin embargo, los valores obtenidos deben considerarse como la autonomía energética mínima, ya que el aporte energético proporcionado por la sección de recolección durante el

funcionamiento del dispositivo TA permitirá aumentar aún más la autonomía del dispositivo. La contribución solar depende de las condiciones ambientales (tipología de fuente, intensidad luminosa, inclinación y orientación de los paneles solares con respecto a la fuente de luz). Sin embargo, las pruebas de campo demuestran que colocando la prenda perpendicularmente al sol, con 52.000 lux de iluminancia y dejando el dispositivo estacionario, la autonomía energética aumenta en aproximadamente un factor 2 (es decir, hasta 6.1 h), en comparación con la ausencia total de cualquier fuente luminosa (por lo general, 3.2 h). Además, en las mismas condiciones, la sección de recolección de energía emplea aproximadamente 3.2 h para cargar completamente la batería de 4600 mAh, antes de conectar el dispositivo TA.

3.2. Software

La estructura del software que se ejecuta en el SoM se muestra en la Figura 5; como se destaca, ésta consta de tres módulos principales: configuración del conjunto de datos, reconocimiento de objetos y posicionamiento de objetos.

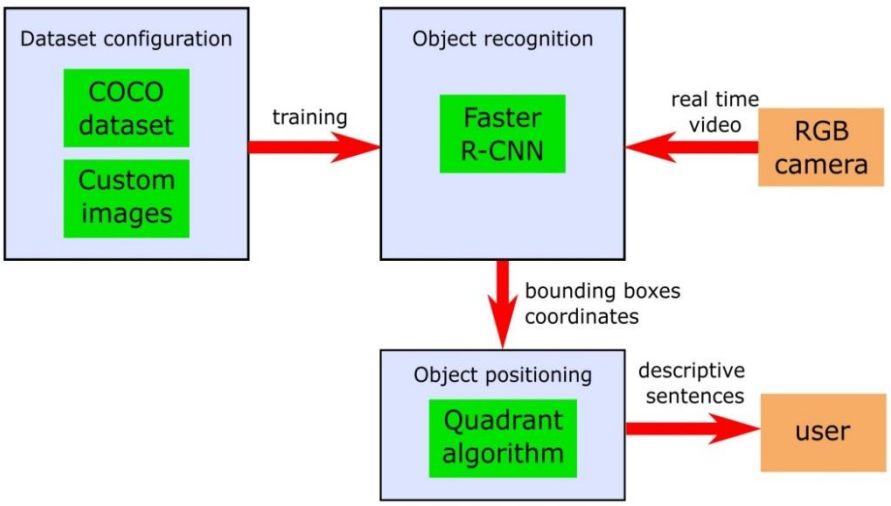


Figura 5. Resumen de la estructura de software.

El módulo de configuración del conjunto de datos abarca la base de datos de Microsoft Common Objects in COntext (COCO) [59], que contiene 91 categorías de objetos comunes y 82 categorías tienen más de 5,000 instancias etiquetadas. En

total, COCO contiene 328,000 imágenes con 2,500,000 instancias etiquetadas. Una característica principal de COCO es que ofrece vistas no canónicas de los objetos (por ejemplo, objetos en el fondo o parcialmente ocultos o en medio del desorden), lo que mejora el rendimiento del reconocimiento. El módulo también puede administrar imágenes personalizadas o imágenes de interés agregadas por el usuario.

El módulo de configuración del conjunto de datos se utiliza para entrenar el módulo de reconocimiento de objetos. Basado en Faster R-CNN, el primero es la columna vertebral del sistema, siendo responsable de detectar y reconocer objetos comunes que se encuentran en escenas y situaciones cotidianas a partir de video en tiempo real. Este módulo genera las coordenadas de los cuadros delimitadores que encierran el objeto. El módulo de orientación detecta la posición del objeto en la imagen a través de un algoritmo de cuadrante cartesiano simple y le permite al usuario conocer la presencia de un objeto y su ubicación a través de frases descriptivas audibles.

El SoM ejecuta un sistema operativo (SO) Linux. Todo el código se implementó empleando el software Python, ejecutándose en Jetson Nano SoM [52,53]. El kit de desarrollador Jetson Nano se configuró de acuerdo con la guía de instalación oficial de Nvidia [60]. En este caso, el archivo JetPack 4.3 y el framework Deepstream se recomiendan para Tensorflow 1.14.0 según la documentación oficial de Nvidia [61]. Se utilizó Python 3.6, siendo la versión más reciente compatible con TensorFlow 1.14.0. Las bibliotecas esenciales utilizadas para respaldar el entrenamiento y la ejecución en el Jetson Nano SoM son las siguientes: Numpy, Pycocotools, PyCuda, OpenCV, Time, Serial y MatPlotLib, PIL, todas en su última versión.

3.2.1. Reconocimiento de objetos

Como se mencionó anteriormente, Faster R-CNN es el último desarrollo de las arquitecturas de la familia R-CNN. Tiene la ventaja de aumentar la eficiencia computacional reduciendo los tiempos de entrenamiento y prueba, además que mejora el rendimiento del reconocimiento de objetos. Su arquitectura se muestra en la Figura 6.

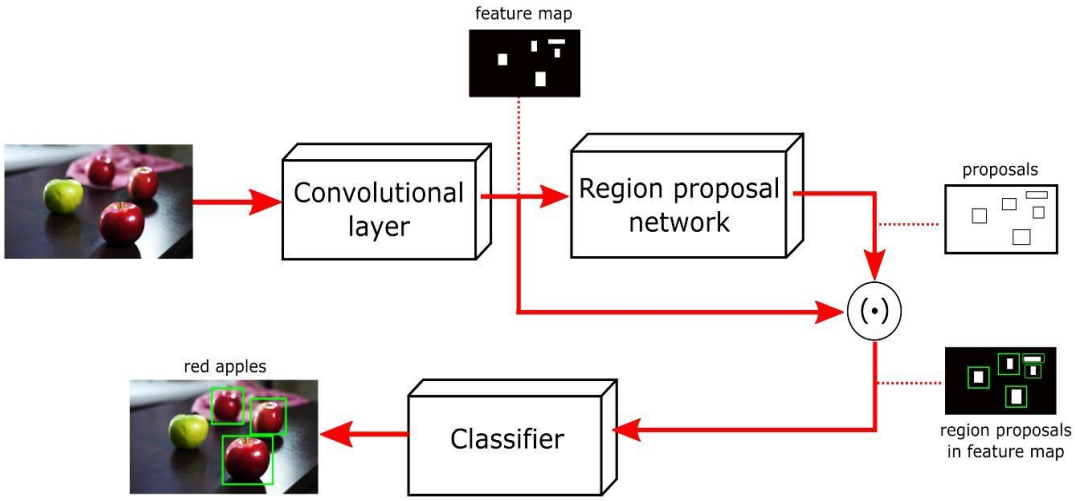


Figura 6. Arquitectura general de Faster R-CNN.

Faster R-CNN consta de tres módulos principales: una capa convolucional, una red para general propuestas de regiones y una capa de clasificación. La capa convolucional es la etapa de extracción de características. Se trata de un conjunto de filtros que se activan cuando detectan características visuales en las imágenes de entrada como bordes, colores, orientaciones específicas, etc. Las salidas de esta etapa son mapas de características. La etapa de generar propuestas de regiones genera ubicaciones de posibles objetos contenidos en los mapas de características. Las propuestas de regiones resultantes se aplican a los mapas de características. Finalmente, la capa de clasificación se utiliza para determinar a qué clase pertenecen los objetos encontrados.

En este trabajo, se utilizó la arquitectura VGG16 para la capa convolucional (Figura 7). VGG16 ha demostrado el mejor rendimiento en tareas de reconocimiento de imágenes [18]. Se ha utilizado anteriormente en el reconocimiento de emociones faciales para predecir la aceptación de los productos alimenticios por parte del consumidor con resultados satisfactorios [62]. VGG16 es una CNN de 16 capas; comprende cinco bloques de filtro que contienen un total de 13 capas de filtro y cinco capas de agrupación. Cada una de las 13 capas de filtro incluye una unidad lineal rectificada (ReLU) para permitir un entrenamiento de redes neuronales más rápido y efectivo. Tres capas completamente conectadas siguen la pila de capas filtradas para aplanar las características de alto nivel en los datos. La función

softmax es la capa final de la arquitectura. Ésta asigna los datos no normalizados a una distribución de probabilidad, que se puede utilizar como entrada de los siguientes módulos.

El módulo "Region Proposal Network" (RPN) (Figura 8) toma el mapa de características como entrada para generar un conjunto de propuestas de objetos rectangulares (denominados cuadros delimitadores). Para generarlos, una máscara se desliza a lo largo del mapa de características. Los valores resultantes se alimentan a dos submódulos paralelos: un clasificador y un regresor. El primero determina la probabilidad p_i de que una propuesta tenga el objeto. El segundo proporciona las coordenadas de píxeles (x, y) de la propuesta.

Finalmente, la capa de clasificación se basa en la arquitectura Fast R-CNN (Figura 9). Su entrada es el mapa de características con las propuestas de la región. Para reducir la cantidad de datos y, por lo tanto, el cálculo a realizar, primero se utiliza una capa de agrupación de RoI (Región de interés); cada valor va a una capa completamente conectada (la de aprendizaje) para detectar las combinaciones no lineales de estas características. Como en la etapa anterior, los valores resultantes se envían a un clasificador y un regresor. El resultado de este módulo es la detección de objetos.

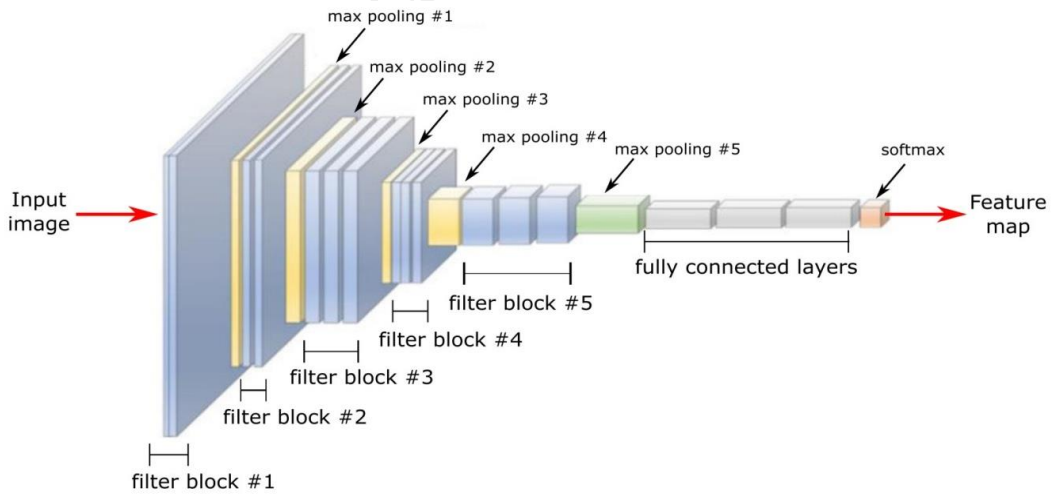


Figura 7. Capa convolucional basada en la arquitectura VGG16.

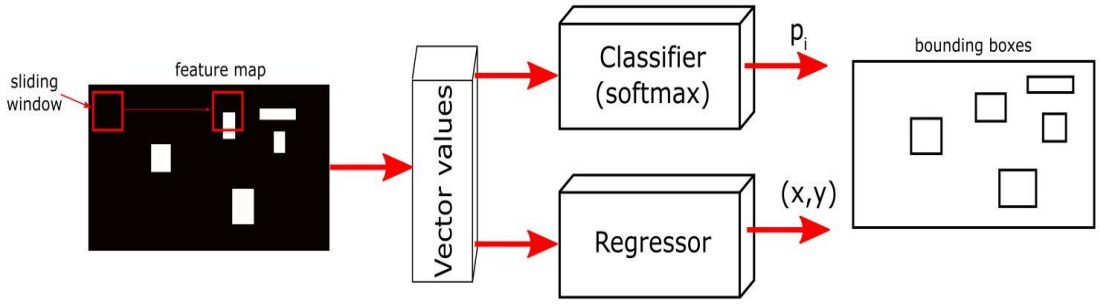


Figura 8. Módulo de regiones propuestas.

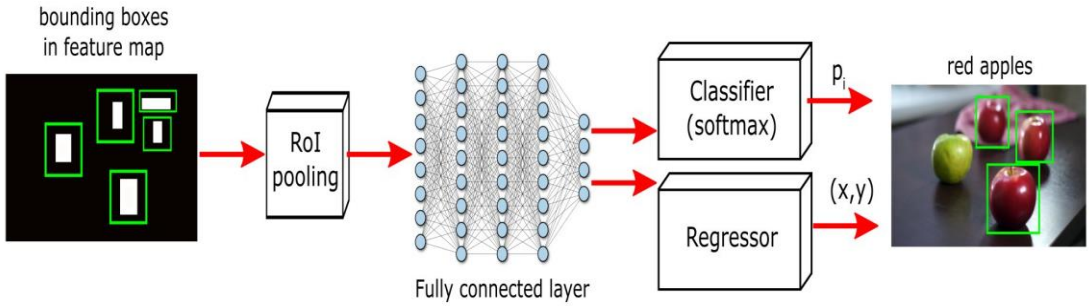
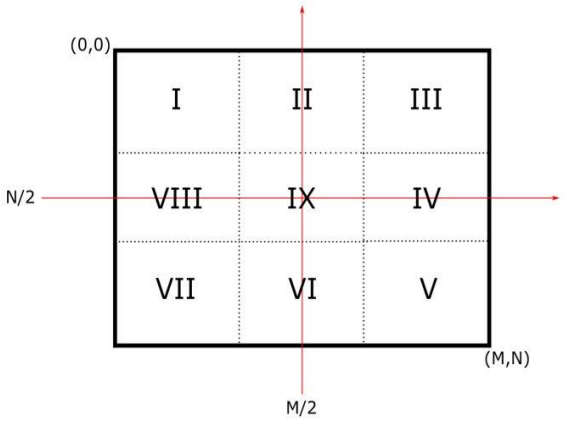


Figura 9. La capa de clasificación basada en la arquitectura Fast R-CNN.

3.2.2. Posicionamiento de objetos

Una vez que se ha detectado un objeto en el FOV de la cámara, es de interés informar al usuario sobre su ubicación. La arquitectura del software abarca un algoritmo de posicionamiento de objetos simple y computacionalmente ligero basado en cuadrantes cartesianos. El concepto se ilustra en la Figura 10. La imagen se divide en cuatro cuadrantes; nueve secciones se definen dentro del espacio cartesiano. Cuando se detecta un objeto, sus coordenadas se pueden posicionar de forma segura en una de estas secciones y se asocia una frase descriptiva audible (ejemplo: “toma de corriente hacia abajo”, “picaporte de puerta al frente”, “bastón blanco hacia abajo”, etc.). De esta forma se notifica a los usuarios de la presencia y ubicación de un objeto para que puedan decidir sobre sus acciones posteriores.



Section	Associated sentence
I	Left up
II	Up
III	Right up
IV	Right
V	Right down
VI	Down
VII	Left down
VIII	Left
IX	In front

Figura 10. Algoritmo de posicionamiento de objetos basado en cuadrantes cartesianos.

Puede haber casos en los que los objetos abarquen dos o más secciones, haciendo ambiguo este enfoque de posicionamiento. Para resolver estos casos, se tomó el centro de masa geométrico (CoM) del cuadro delimitador que encierra el objeto como un indicador confiable para decidir sobre la sección de posicionamiento.

3.2.3. Implementación

La Figura 11 muestra el diagrama de casos de uso del sistema que resume su funcionalidad. El primer paso consiste en que la cámara obtenga una imagen. En paralelo, el acelerómetro integrado al SoM determina si el usuario está en movimiento. Si es así, el Jetson Nano SoM se configura en modo de consumo de energía normal (10 W); de lo contrario, permanecerá en el modo de bajo consumo (5 W). Una vez que se ha capturado una imagen, el Faster R-CNN la analiza en busca de objetos definidos en la base de datos. En caso de que una o más detecciones sean positivas, se dibujarán rectángulos. A continuación, el algoritmo de cuadrantes cartesiano coloca el(los) objeto(s) en el marco. Finalmente, el altavoz transmite la información al usuario.

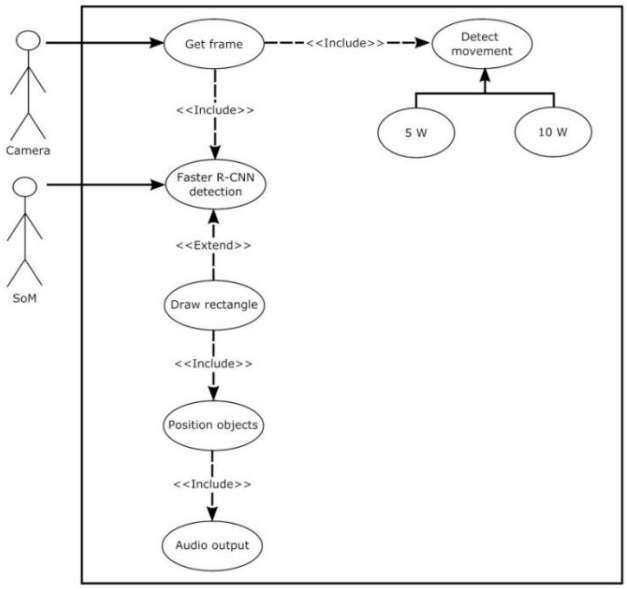


Figura 11. El diagrama de casos para el dispositivo TA.

El algoritmo 1 muestra el pseudocódigo de operación del software en el SoM que integra los tres módulos descritos anteriormente.

Algoritmo 1: El pseudocódigo del dispositivo TA para la detección y el posicionamiento de objetos.

```
While camera is active
 | Get frame;
 | Run Object detection for frame;
 | If object detected==true
  | Obj_label = object category;
  | Get (x,y) coordinates of bounding box;
  | Run Object positioning for object detected;
  | Obj_pos = section where the object is located;
  | Pos_string = string associated to Obj_pos;
  | Audio = concatenate(Obj_label, Pos_string);
  | Output Audio;
  | Clear (x,y) coordinates;
  | end
 | end
```

4. Resultados experimentales

4.1. Entrenamiento del sistema de detección de objetos

Como se mencionó en la Sección 3.2, el conjunto de datos COCO se utilizó para entrenar el módulo de detección de objetos. Se usó un modelo gráfico de inferencia congelado, el cual es una arquitectura Faster R-CNN previamente entrenada por COCO. Este modelo sirve como base del módulo de configuración del conjunto de datos (ver la Figura 5). Los archivos del modelo se pueden descargar de [63] y contienen las 91 categorías de objetos de COCO (ver Tabla 3). También se pueden agregar objetos personalizados al módulo de configuración del conjunto de datos. Para este proyecto, se consideraron cuatro categorías adicionales: tomacorriente, picaporte, interruptor y bastón blanco. Se incorporaron un total de 400 imágenes (aproximadamente 100 por categoría) a la base de datos general.

El entrenamiento de estos cuatro objetos adicionales se llevó a cabo externamente utilizando una computadora de escritorio con las siguientes especificaciones de hardware: 8 Gb de RAM, procesador AMD Ryzen 5 2500U y NVIDIA GTX 1050 con 4 GB de VRAM. Utilizando este equipo, se realizaron un total de 200,000 iteraciones en aproximadamente 20 horas. Los parámetros de configuración de entrenamiento del Faster R-CNN se muestran en la Tabla 4, con una descripción detallada de su significado. Estos valores se utilizan normalmente para entrenar este tipo de estructuras basadas en el aprendizaje profundo de forma óptima [17].

Posteriormente, tanto los archivos COCO como los personalizados se cargan en el módulo SoM. Los archivos no se pueden modificar una vez en el SoM, pues son solo para ejecución. Esto representa una ventaja: aún si la batería del SoM se agota, los archivos entrenados permanecen listos para usarse una vez que se restablece la energía. La Tabla 3 resume las capacidades de detección de objetos del dispositivo de TA propuesto.

Tabla 3. Objetos detectados por el dispositivo TA.

Imágenes en la base de datos COCO					**Objetos persona-lizados**
Manzana	Zanahoria	Jirafa	Persona	Maletín	Picaporte
Mochila	Gato	Peine	Pizza	Tabla de Surf	Enchufe
Banana	Celular	Secador	Plato	Oso de peluche	Interruptor
Bate	Silla	Mochila de mano	Planta	Raqueta de tenis	Bastón
Guante de baseball	Reloj	Sombrero	Refrigerador	Corbata	
Oso	Sofá	Caballo	Control remoto	Tostadora	
Cama	Vaca	Hot dog	Sandwich	Retrete	
Banca	Taza	Teclado	Tijeras	Cepillo de dientes	
Bicicleta	Escritorio	Papalote	Oveja	Semáforo	
Pájaro	Mesa de comida	Cuchillo	Zapato	Tren	
Licuadora	Perro	Laptop	Pileta	Camión	
Bote	Dona	Microondas	Skate	Tv	
Libro	Puerta	Espejo	Esquí	Sombrilla	
Botella	Elefante	Moto	snowboard	Florero	
Bol	Lentes	Ratón	Cuchara	Ventana	
Brócoli	Boca de incendio	Naranja	Pelota	Copa de vino	
Camión de personal	Tenedor	Horno	Señal de Stop	Cebra	
Pastel	Frisbee	Parquímetro	Señal	Maleta	
Auto					

Tabla 4. Parámetros de configuración de entrenamiento utilizados para el Faster R-CNN.

Parámetro de entrenamiento	Valor	Descripción
Total trained classes	95	El número total de objetos (91 del conjunto de datos COCO más 4 personalizados).
Dimensions (pixels)	1024 x 768	Resolución de las imágenes en fase de entrenamiento.
Batch size (images)	12	El número de muestras procesadas antes de que se actualice el modelo.
Optimizer	Adaptive Moment Estimation (Adam)	Los optimizadores son métodos que se utilizan para configurar los atributos de la red neuronal como los pesos y la tasa de aprendizaje con el fin de reducir las pérdidas.
Learning rate (first 60 k mini-batches)	0.0002	La tasa de aprendizaje es un hiperparámetro que controla la respuesta del modelo ante el error estimado cada vez que se actualizan los pesos del modelo.
Learning rate (60 k–120 k mini-batches)	0.00002	
Weight decay	0.0005	Weight decay permite que la red neuronal disminuya su complejidad y reduzca el tiempo de entrenamiento sin penalizar la precisión de los detectores.
Intersection over Union (IoU)	0.55	IoU es una métrica de evaluación que se utiliza para medir la precisión de un detector de objetos en un conjunto de datos en particular.
Dropout	FALSE	Dropout es un método de regularización que se aproxima al entrenamiento de una gran cantidad de redes neuronales con diferentes arquitecturas en paralelo.
Shuffle	TRUE	Shuffle se usa para mezclar el conjunto de datos personalizado, por lo que evita que la red neuronal lo memorice.
Max detections per class	100	Es el número máximo de detección de un objeto (clase) en un cuadro.
Max total detections	300	Detecciones máximas permitidas en un cuadro.

4.2. Reconocimiento de objetos

Una de las ventajas de utilizar estructuras basadas en el aprendizaje profundo es que funcionan en una amplia gama de condiciones: nivel de iluminación, entornos interiores / exteriores, diferentes ángulos de visión, oclusiones parciales, diferentes propiedades del objeto (dimensiones, colores, texturas, etc.), entre otros. Por lo tanto, no es necesario tener un ambiente controlado para probar el sistema y obtener buenos resultados.

Las Figuras 12 y 13 muestran algunos resultados en el reconocimiento de un solo objeto y de múltiples objetos realizado con el enfoque propuesto. En este trabajo experimental, se han considerado escenas tanto en interiores como en exteriores que presentan diferentes niveles de iluminación, ángulos de visión de objetos, oclusiones parciales de objetos y propiedades de los objetos. Nótese que el reconocimiento de objetos tanto los incluidos en COCO como los personalizados se realiza con alta precisión y solidez.

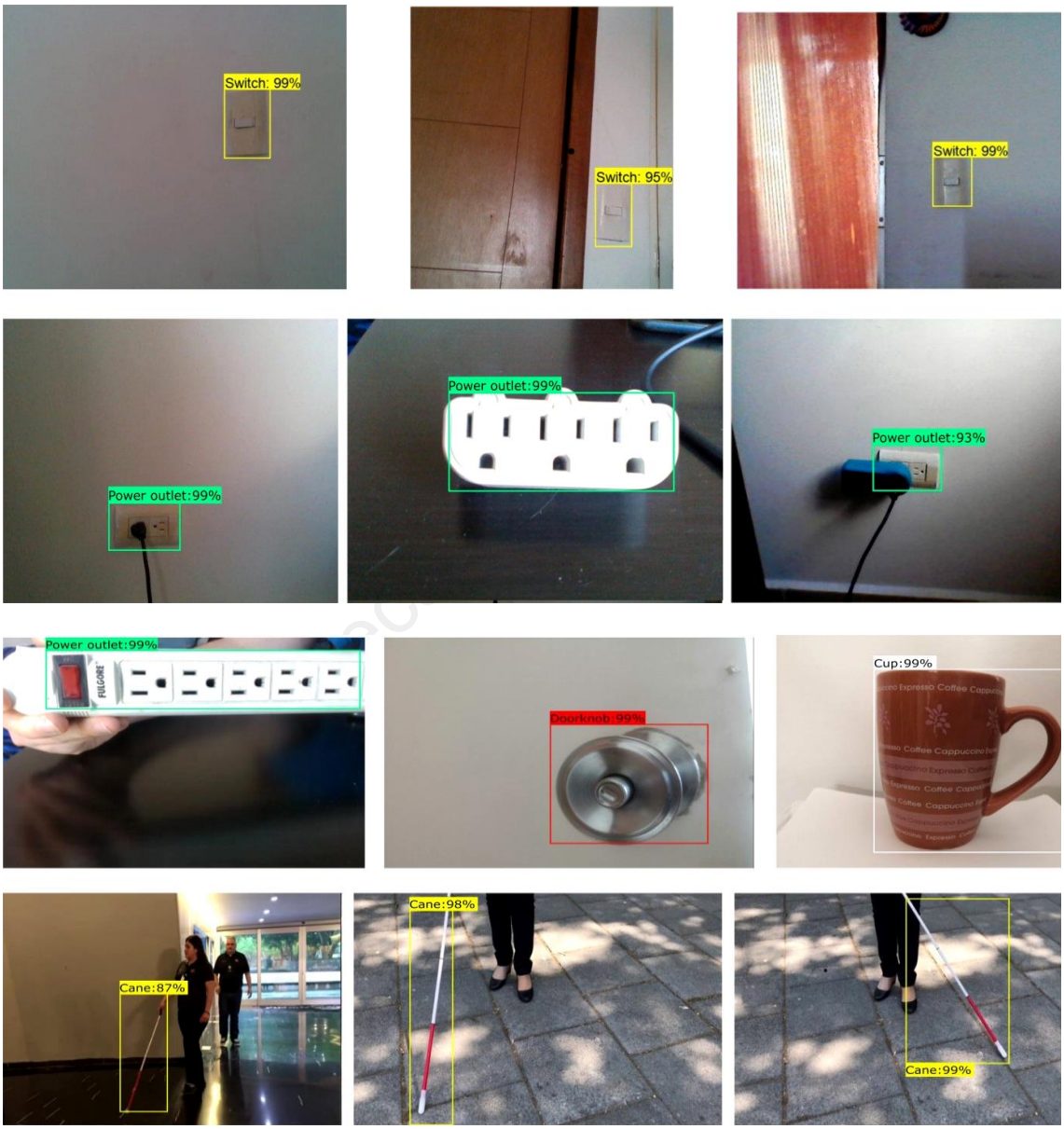


Figura 12. Ejemplos de detección de un solo objeto con el dispositivo portátil. Objetos personalizados: interruptor, toma corriente, picaporte y bastón blanco. Clases de COCO: taza.

Los cuadros delimitadores que encierran los objetos reconocidos incluyen un valor de confianza, es decir, la probabilidad (en %) de que el cuadro delimitador contenga un objeto conocido. La Figura 14 muestra un ejemplo de valores de confianza decrecientes de un objeto debido a su oclusión progresiva (de izquierda a derecha en la Figura 14). Observe en la imagen central que el dedo dentro del área del cuadro delimitador disminuye el valor de confianza del 93% al 87%. Al igual que la visión humana, existe un límite en el que el dispositivo de TA es capaz de detectar un objeto.

Como se mencionó anteriormente, el reconocimiento de objetos con aprendizaje profundo es robusto para muchos parámetros de la escena, uno de ellos es el ángulo de visión del objeto, particularmente relevante porque la cámara está montada sobre el armazón de los lentes del usuario. A medida que la cabeza se mueve, independientemente de que el usuario esté estático o en movimiento, es posible que los objetos capturados en el marco no estén alineados con la línea de base del suelo. La Figura 15 muestra un conjunto de ejemplos de objetos desalineados con la línea de base del suelo; note que el reconocimiento es independiente del ángulo de visión de la cámara.

El procesamiento de video en tiempo real generalmente conduce a algunos falsos positivos que aparecen solo en uno o dos cuadros durante una ventana de tiempo. La principal causa de los falsos positivos es la variación de la iluminación [64]. Para filtrar los casos inciertos y reducir así los falsos positivos, se estableció un umbral de confianza del 55%. Por tanto, los objetos por debajo de este valor de probabilidad no se tendrán en cuenta [65,66].

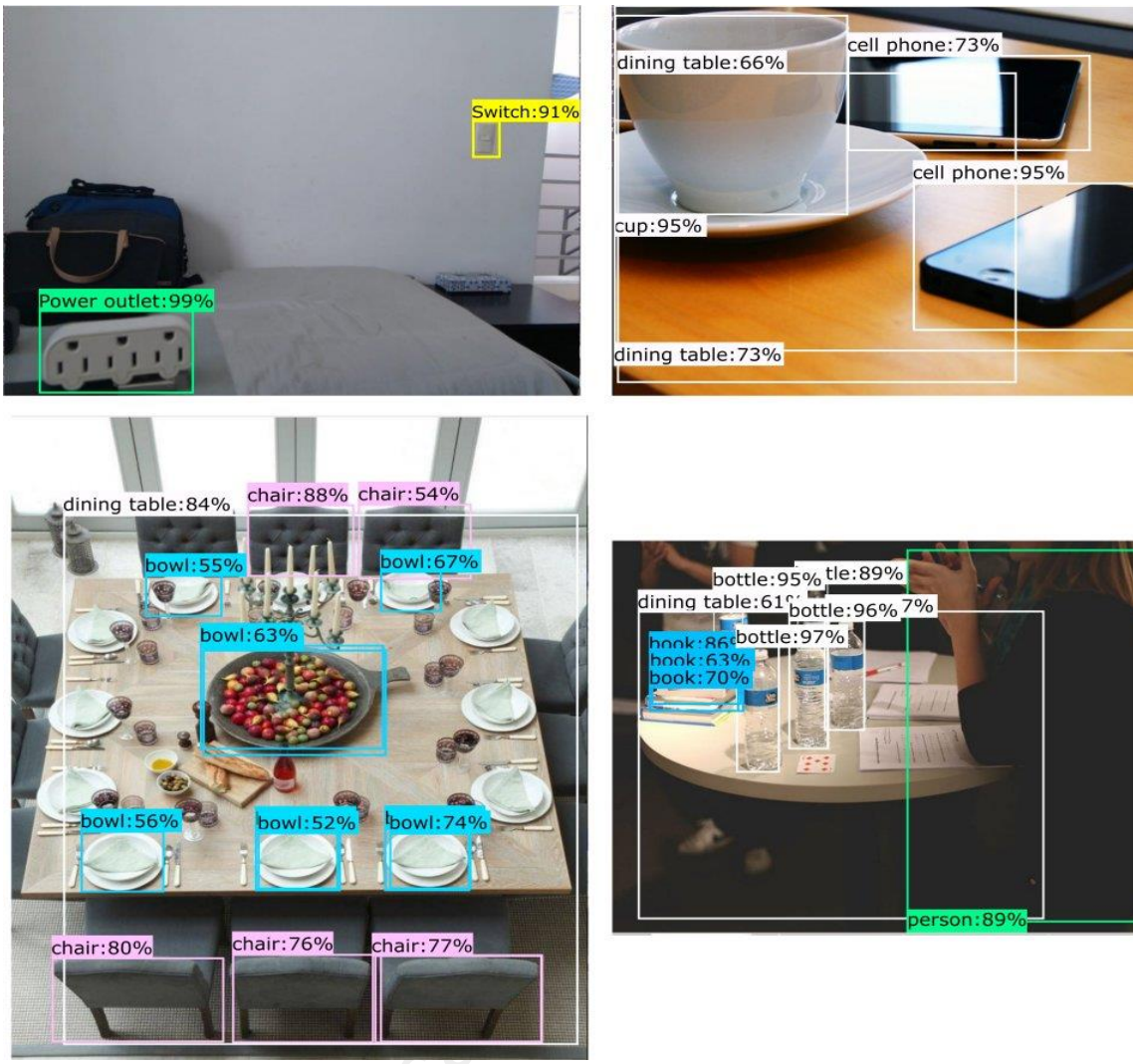


Figura 13. Ejemplo de detección de múltiples objetos. Clases personalizadas: interruptor y toma corriente. Clases COCO: taza, teléfono celular, mesa, silla, tazón, libro, botella y persona.

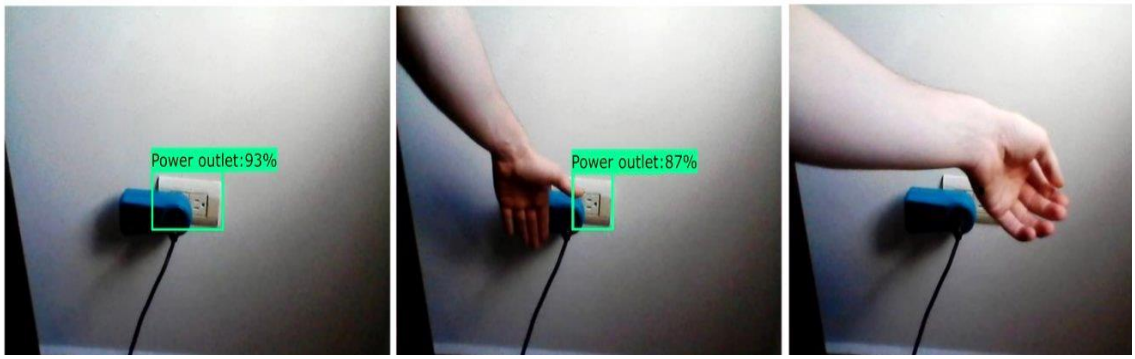


Figura 14. Ejemplo de disminución progresiva del valor de confianza debido a la oclusión del objeto (de izquierda a derecha).

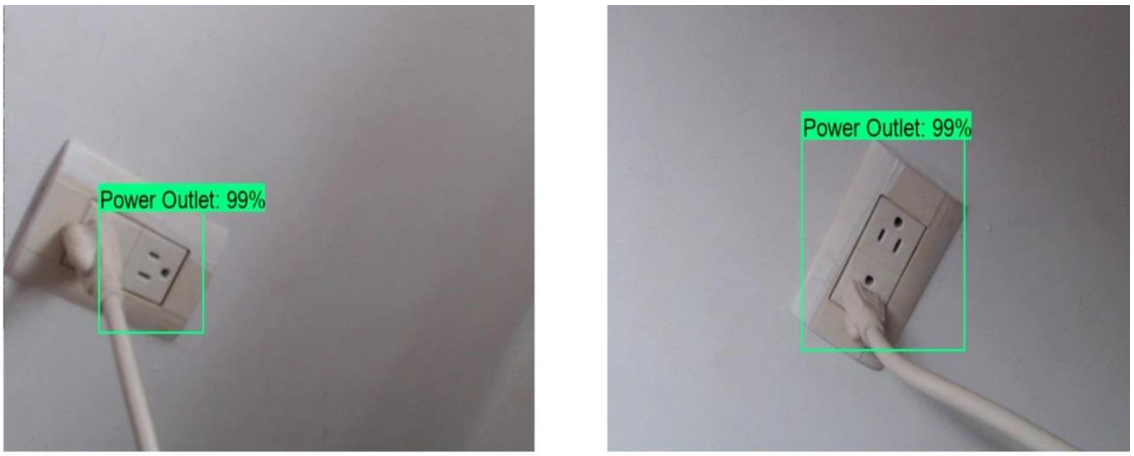


Figura 15. Ejemplo de reconocimiento de objetos con diferentes ángulos de visión.

4.3. Posicionamiento de objetos

La Figura 16 muestra un ejemplo representativo del desempeño del módulo de posicionamiento de objetos. El algoritmo de cuadrantes cartesiano ubica el toma de corriente en la Sección VII mientras que el interruptor en la Sección IV (ver en la Figura 10). Note la simplicidad y eficacia del algoritmo.

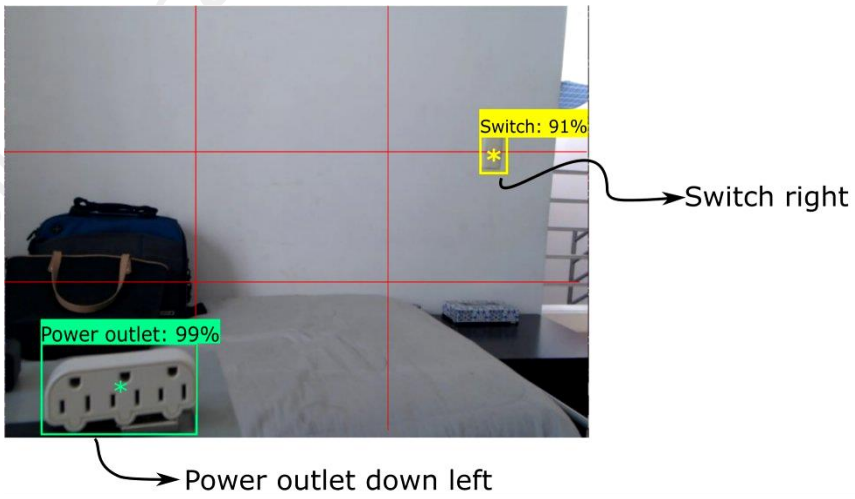


Figura 16. Ejemplo de posicionamiento de objetos basado en cuadrantes cartesianos. El símbolo * representa el CoM del cuadro delimitador. Éste se toma como un factor de decisión y es especialmente útil en los casos en que los objetos abarcan más de una sección.

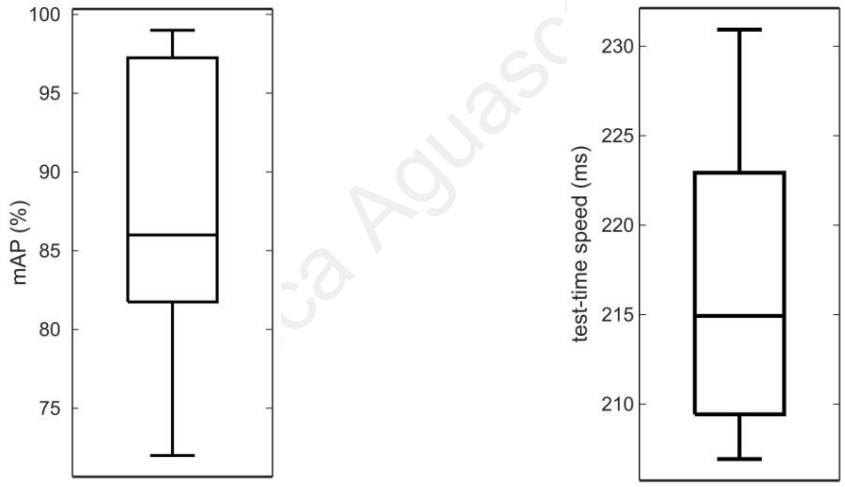
A medida que el usuario mueve libremente la cabeza o se acerca al objeto, la posición relativa del elemento puede cambiar en la imagen, cambiando así la sección asignada por el algoritmo. Esto no plantea ningún problema, ya que las frases descriptivas sonoras se actualizan constantemente. Es importante considerar que este módulo no pretende proporcionar información de posicionamiento que pueda permitir una aproximación precisa hacia los objetos, sino ofrecer a los usuarios una visión general de dónde se encuentran tomando como referencia su mirada (la cámara sobre los lentes se mueve según la mirada del usuario).

4.4. Métricas de rendimiento

En esta subsección, se detallan las siguientes métricas de rendimiento que son de interés para nuestro dispositivo TA. Para la detección general de objetos, la precisión de reconocimiento promedio (mAP) y el tiempo promedio de procesamiento. Para la clasificación de objetos: las funciones de pérdida. Finalmente, para el módulo de conjunto de datos: el valor de confianza en función de las imágenes utilizadas en el entrenamiento.

Para determinar correctamente el mAP y el tiempo promedio de procesamiento (es decir, el reconocimiento de objetos y su posicionamiento), el dispositivo TA se sometió a distintas condiciones de funcionamiento. Un total de 25 objetos diferentes (los cuatro objetos personalizados y los 21 de COCO) fueron capturados por la cámara en diferentes vistas, ángulos y distancias (hasta 5 m de distancia del usuario), niveles de iluminación, así como en diferentes condiciones ambientales (exterior e interior de una casa). Se procesó un conjunto de 7,500 imágenes en tiempo real. La Figura 17 muestra un análisis de diagrama de caja que resume los resultados obtenidos. El mAP se puede estimar en 86% (Figura 17a), mientras que el tiempo promedio de procesamiento en 215 ms (que corresponde a una tasa de actualización para la información de 4,65 Hz), específicamente para el modo de funcionamiento (modo 0) con mayor frecuencia de reloj (Figura 17b). Estos números nos permiten confirmar que se puede realizar un reconocimiento de objetos en tiempo real preciso y confiable con el enfoque propuesto.

La Figura 18 muestra un conjunto de funciones de pérdida que representan la inexactitud de las predicciones durante el proceso de clasificación de la capa de clasificación (Figura 9). La figura 18(a) muestra la función de pérdida para la clasificación inicial, es decir, el reconocimiento de objetos únicamente. Note que la inexactitud disminuye rápidamente con el número de iteraciones. Las 200,000 iteraciones realizadas para las cuatro imágenes personalizadas (Sección 4.1) nos permiten estimar una inexactitud de tan solo el 2.5%; con respecto a esto, en referencia a la Figura 18a, se debe tener en cuenta que para el valor de 200K iteraciones (eje X), se obtuvo un valor de inexactitud de 0.025, es decir 2.5%. Para mayor claridad con referencia a las Figuras 18, el valor de inexactitud (es decir, "Función de pérdida") proviene de 1 (es decir, 100%) durante las primeras iteraciones, pero disminuye rápidamente, a medida que la red está aprendiendo y aumenta el número de iteraciones (eje X).



(a) (b)

Figura 17. Análisis de diagrama de caja para (a) el mAP y (b) el tiempo promedio de procesamiento

De manera similar, la Figura 1 (b) muestra la función de pérdida para la localización, es decir, los cuadros delimitadores que encierran los objetos. Se puede esperar una inexactitud del 4% en el trazado de los cuadros delimitadores. La Figura 18(c) muestra la función de pérdida para toda la capa de clasificación, es decir, el reconocimiento y la localización de objetos. Se observa un error del 6% sobre el

número de iteraciones. En conclusión, las funciones de pérdida garantizan una detección de objetos precisa y fiable.

La Figura 19 muestra la evolución del valor de confianza en función del número de imágenes utilizadas para el entrenamiento en el módulo de configuración del conjunto de datos. Se observa un comportamiento logarítmico. Note que las 40 imágenes garantizan el umbral mínimo de detección del 55% y se pueden utilizar para un entrenamiento rápido. Las 400 imágenes utilizadas para entrenar los objetos personalizados logran una precisión del 89%.

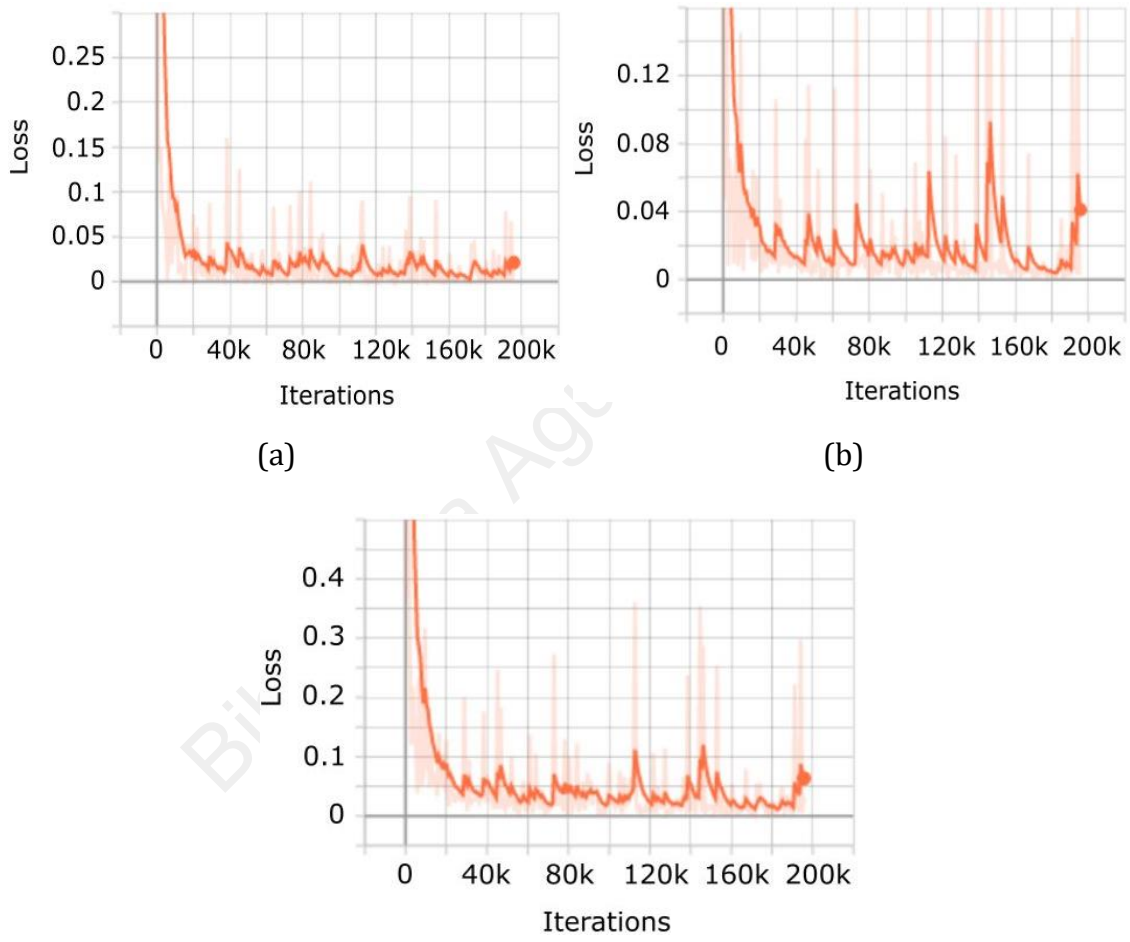


Figura 18. Funciones de pérdida (%) para la capa de clasificación en función del número de iteraciones: (a) clasificación de objetos, (b) localización y (c) clasificación general de capas.

Figura 19. El valor de confianza en función de las imágenes utilizadas para el entrenamiento.

5. Discusión de resultados

Los resultados experimentales obtenidos confirman un reconocimiento de objetos en tiempo real con una precisión media (mAP) del 86% y un tiempo promedio de procesamiento de 215 ms en el caso de la modalidad de funcionamiento en modo 0 con reloj de alta frecuencia (valor que aumenta hasta 360 ms en caso de la habilitación del modo 1 de baja velocidad). La función de pérdida de la clasificación general de objetos muestra que el proceso de entrenamiento se realizó correctamente y que se pueden esperar inexactitudes de clasificación del orden del 6%.

Una comparación directa con otros sistemas descritos en la literatura, dirigidos a la asistencia de las personas con DV, no es evidente ya que la mayoría de ellos no reportan sus mAPs [29, 30, 32]. Otros trabajos de reconocimiento de objetos [67-69] para diferentes aplicaciones informan mAPs entre 70% y 90% explorando otros enfoques de aprendizaje profundo como SSD (Single Shot MultiBox Detector), YOLO (You Only Look Once) y R-FCN (Región-basadas en Redes Totalmente Convolucionales) con el uso de otras bases de datos de imágenes para el entrenamiento, como PASCAL VOC (Clases de Objetos Visuales), SUN (Entendimiento de Escena) y KITTI (Instituto Tecnológico de Karlsruhe e Instituto Tecnológico Toyota).

No obstante, nuestro mAP coincide con los descritos en [70] y [71]. En el primero, se obtuvo una mAP del 81.09% para la detección de vehículos en movimiento. En el segundo trabajo, los autores reportan 86.67% de mAP para la detección de objetos en imágenes deportivas. Ambos exploran la misma dupla de este trabajo: Faster R-CNN y COCO; sin embargo, sus plataformas informáticas no son una solución compacta como el SoM Jetson Nano que se empleó en este trabajo. El estudio del valor de confianza confirma evidentemente que cuantas más imágenes se involucren en el entrenamiento de la red, mejor. Para la arquitectura de red del dispositivo TA, se pueden lograr entrenamientos completos que garantizan tasas de reconocimiento del 89% con 400 imágenes, mientras que un entrenamiento rápido realizado con solo 40 imágenes puede garantizar el umbral mínimo de confianza del 55%.

El trabajo futuro explorará otras técnicas de visión por computadora, como el análisis de múltiples escalas, la segmentación de imágenes, la resolución múltiple, las pirámides, etc. para mejorar aún más el mAP actual. Estos métodos ciertamente requerirán otros tipos de hardware para realizar la detección en tiempo real. Se realizará una evaluación integral que involucra el mAP, tiempo de computación, capacidad de uso del hardware y costo para determinar la mejor opción.

El sistema propuesto está listo para la evaluación de usuarios. El trabajo actual se centra en el diseño de las pruebas de usuario. Su objetivo será evaluar cómo hacen uso de la información proporcionada por el dispositivo. En particular, buscamos determinar la frecuencia de actualización de la retroalimentación audible adecuada, es decir, con qué frecuencia es pertinente transmitir las oraciones descriptivas al usuario. Trabajos previos en TA para personas con DV [5,7,10] han proporcionado ideas interesantes sobre cómo se debe transmitir la información a los usuarios: debe ser simple y no continua, porque seguramente abrumará y fatigará cognitivamente a los usuarios que finalmente podrían rechazar el dispositivo, independientemente de sus avances tecnológicos y beneficios. Además, investigaremos la pertinencia del módulo de posicionamiento de objetos. Estamos interesados en determinar si la información general del espacio 2D es lo suficientemente útil para que los usuarios interactúen con el entorno. Una información más detallada que proporcione distancia de un objeto puede ser demasiado demandante a nivel cognitivo. Además, la información en 3D implicaría el uso de dispositivos adicionales para determinar

la profundidad y la proximidad del usuario al objeto, como cámaras y sensores IMU adicionales.

En la Tabla 5, se compara el dispositivo de TA propuesto en esta tesis con otros sistemas de reconocimiento visual, presentados anteriormente en la Sección 2; en concreto, las principales características consideradas para la comparación son el mAP, el tiempo promedio de procesamiento, la complejidad de la solución propuesta, ambos teniendo en cuenta la carga computacional requerida para la detección de objetos y los recursos para entrenar el algoritmo de reconocimiento, y finalmente el costo de la solución propuesta.

En particular, nuestra solución obtiene un mAP satisfactorio (86%) y tiempo de procesamiento (215 ms) en comparación con otros sistemas similares que requieren procesadores de alto rendimiento y alta frecuencia difícilmente aplicables para un sistema portátil. Por ejemplo, el sistema en [26] admite una velocidad de 30 fps con el uso de una computadora de alto rendimiento con un procesador Intel Core i7 de 2,4 GHz y 32 GB de RAM; estas condiciones hacen que la integración del sistema en un dispositivo portátil sea muy difícil, lo que implica altos costos de implementación.

Tabla 5. Comparación entre el sistema de detección de objetos propuesto y otros trabajos similares reportados en la literatura.

Solución	mAP [%]	Tiempo medio de cálculo [ms]	Complejidad	Costo
L.B. Neto [24]	94	2.4	Alto	Alto
S. Lu [38]	88	22	Alto	Alto
U. Malūkas [40]	96	105	Alto	Alto
K. Jayakanth [41]	100	451	Medio-Alto	Alto
Our system	86	215	Bajo	Bajo

6. Conclusiones

Esta tesis ha presentado un dispositivo innovador de TA portable cuyo propósito es asistir a las personas con DV a encontrar objetos de uso diario en el entorno cercano. Se han analizado todos los aspectos del sistema: (i) un análisis

extenso del estado del arte con una comparación con dispositivos y técnicas similares que permitió apreciar la precisión y rentabilidad de la solución propuesta, (ii) rendimiento de la técnica de aprendizaje profundo adoptada (Faster R-CNN), (iii) posibilidad de extender la autonomía energética del sistema mediante el uso de un sistema de recolección de energía solar portátil para suministrar energía al dispositivo TA, (iv) el rendimiento del dispositivo TA relacionado con reconocimiento de objetos, lo que demuestra que el sistema implementado es robusto para muchos parámetros de la escena y puede detectar con alta precisión las imágenes en tiempo real de una biblioteca extendida.

El sistema está compuesto por una cámara en miniatura de bajo costo y un sistema de módulos (SoM). La cámara se coloca en el armazón de los lentes del usuario y captura video en tiempo real del entorno circundante. El SoM se puede colocar al nivel de la cintura y procesa las imágenes enviadas por la cámara. La batería y el módulo electrónico del sistema portátil de recolección de energía solar, útil para aumentar la autonomía del dispositivo TA diseñado, se han colocado en bolsillos internos de una chaqueta que viste el usuario. Los algoritmos que se ejecutan en el SoM son capaces de detectar objetos que se encuentran en escenas cotidianas y colocarlos en el espacio cartesiano a través de una metodología que divide la imagen analizada en 9 secciones. La retroalimentación para el usuario consiste en frases descriptivas audibles que involucran el objeto (s) detectado y su (s) posición (es) en el campo de visión de la cámara.

La arquitectura del software presenta tres módulos principales: configuración del conjunto de datos, detección de objetos y posicionamiento de objetos. El módulo de configuración del conjunto de datos contiene un Faster R-CNN previamente entrenado que abarca los 91 objetos del conjunto de datos COCO y también es capaz de alojar imágenes personalizadas. El módulo de detección de objetos consiste en un Faster R-CNN que procesa el video en tiempo real proveniente de la cámara. Éste se encarga de detectar y reconocer los objetos para personas con DV. Finalmente, el módulo de posicionamiento de objetos utiliza un algoritmo de cuadrante cartesiano para posicionar los objetos encontrados y transmite mensajes audibles apropiados al usuario.

Los resultados finales obtenidos con el dispositivo propuesto y presentado en esta tesis son muy alentadores; de hecho, con un entrenamiento completo realizado

en 400 imágenes, es posible alcanzar el 89% de precisión en el reconocimiento de objetos, mientras que con un entrenamiento reducido realizado con solo 40 imágenes, es posible alcanzar un valor de precisión de reconocimiento del 55%.

Finalmente, la visión de este proyecto no se limita a implementar un prototipo de laboratorio. Prevemos una transferencia tecnológica que pueda ayudar a la población con DV, considerando también la economía propuesta del dispositivo (costo total de solo $200 USD), la autonomía energética prolongada y la facilidad de uso. El trabajo futuro incluirá la implementación de un sitio web donde los usuarios puedan descargar los archivos Faster R-CNN ya entrenados. Diferentes categorías como electrodomésticos, equipo de oficina, productos de limpieza, ropa, entre muchas otras estarán listas para su descarga e instalación en el dispositivo TA.

7. Referencias

1. World Health Organization, Fact sheet on Blindness and Vision Impairment (October 2019), https://www.who.int/en/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed 02 August 2020)
2. Bourne, R.R; Flaxman, S.R., Braithwaite, T., Cicinelli, M.V., et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. The Lancet Global Health, 2017, 5, 888-897.
3. Velazquez, R. Wearable Assistive Devices for the Blind. In Wearable and Autonomous Biomedical Devices and Systems for Smart Environment: Issues and Characterization; Lay-Ekuakille, A., Mukhopadhyay, S.C., Eds.; LNEE, 75; Springer: Berlin/Heidelberg, Germany, 2010; pp. 331–349. ISBN 978-3-642-15687-8.
4. National Academies of Sciences, Engineering, and Medicine. Making Eye Health a Population Health Imperative: Vision for Tomorrow. Washington, DC: The National Academies Press. 2016.
5. Velazquez, R.; Fontaine, E.; Pissaloux, E. Coding the Environment in Tactile Maps for Real-Time Guidance of the Visually Impaired. In Proceedings of the IEEE International Symposium on MicroNanoMechanical and Human Science, Nagoya, Japan, 5-8 Nov. 2006; pp. 1-6.

6. Bologna, G.; Deville, B.; Diego-Gomez, J.; Pun T. Toward Local and Global Perception Modules for Vision Substitution. *Neurocomputing*. **2017**, 74(8), 1182–1190.
7. Velazquez, R.; Pissaloux, E.; Rodrigo, P.; Carrasco, M.; Giannoccaro, N.I.; Lay-Ekuakille, A. An Outdoor Navigation System for Blind Pedestrians Using GPS and Tactile-Foot Feedback. *Appl. Sci.* **2018**, 8, 578.
8. Real, S.; Araujo, A. Navigation Systems for the Blind and Visually Impaired: Past Work, Challenges, and Open Problems. *Sensors (Basel).* **2019**, 19, 3404.
9. Bhowmick, A.; Hazarika, S. M.; An insight into assistive technology for the visually impairedand blind people: state-of-the-art and future trends. *Journal on Multimodal User Interfaces* **2017**, 11(2), 1-24 DOI 10.1007/s12193-016-0235-6.
10. Velazquez, R.; Hernandez, H.; Preza, E. A Portable Piezoelectric Tactile Terminal for Braille Readers. *Applied Bionics and Biomechanics.* **2012**, 9(1), 45-60.
11. Neto, R.; Fonseca, N. Camera Reading for Blind People. *Procedia Technology.* **2014**, 16, 1200-1209.
12. Oproescu, M.; Iana, G.; Bizon, N.; Novac, O.C.; Novac, M.C. Software and Hardware Solutions for Using the Keyboards by Blind People. In Proceedings of the International Conference on Engineering of Modern Electric Systems, Oradea, Romania, 13-14 June 2019; pp. 25-28.
13. Watanabe, T.; Kaga, H.; Shinkai, S. Comparison of Onscreen Text Entry Methods when Using a Screen Reader. *IEICE Transactions on Information and Systems*. **2018**, E101.D(2), 455-461.
14. Rawat, W.; Wang, Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*. **2017**, 29(9), 2352-2449.
15. Jmour, N.; Zayen, S.; Abdelkrim, A. Convolutional Neural Networks for Image Classification. In Proceedings of the International Conference on Advanced Systems and Electric Technologies, Hammamet, Tunisia, 22-25 March 2018; pp. 397-402.
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23-28 June 2014; pp. 580-587.

17. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7-13 Dec. 2015; pp. 1440-1448.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **2017**, 39(6), 1137-1149.
19. Meshram, V.V.; Patil, K.; Meshram, V.A.; Shu, F.C. An Astute Assistive Device for Mobility and Object Recognition for Visually Impaired People. *IEEE Trans. Hum.-Mach. Syst.* **2019**, *49*, 449–460, doi:10.1109/THMS.2019.2931745.
20. Krishnan, A.; Deepakraj, G.; Nishanth, N.; Anandkumar, K.M. Autonomous walking stick for the blind using echolocation and image processing. In Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I); IEEE: Noida, India, 2016; pp. 13–16.
21. Cardin, S.; Thalmann, D.; Vexo, F. A wearable system for mobility improvement of visually impaired people. *Vis. Comput.* **2007**, *23*, 109–118, doi:10.1007/s00371-006-0032-4.
22. Chen, S.; Yao, D.; Cao, H.; Shen, C. A Novel Approach to Wearable Image Recognition Systems to Aid Visually Impaired People. *Appl. Sci.* **2019**, *9*, 1–20, doi:10.3390/app9163350.
23. Li, B.; Zhang, X.; Munoz, J.P.; Xiao, J.; Rong, X.; Tian, Y. Assisting blind people to avoid obstacles: An wearable obstacle stereo feedback system based on 3D detection. In Proceedings of the 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO); Zhuhai, China, 2015; pp. 2307–2311.
24. Neto, L.B.; Grijalva, F.; Maike, V.R.M.L.; Martini, L.C.; Florencio, D.; Baranauskas, M.C.C.; Rocha, A.; Goldenstein, S. A Kinect-Based Wearable Face Recognition System to Aid Visually Impaired Users. *IEEE Trans. Hum.-Mach. Syst.* **2017**, *47*, 52–64, doi:10.1109/THMS.2016.2604367.
25. Katzschmann, R.K.; Araki, B.; Rus, D. Safe Local Navigation for Visually Impaired Users With a Time-of-Flight and Haptic Feedback Device. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 583–593, doi:10.1109/TNSRE.2018.2800665.
26. Chen, T.; Ravindranath, L.; Deng, S.; Bahl, P.; Balakrishnan, H. Glimpse: Continuous, Real-Time Object Recognition on Mobile Devices. In Proceedings of

the 13th ACM Conference on Embedded Networked Sensor Systems, New York, NY, USA, November 2015; pp. 155–168.

27. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. *International Journal of Computer Vision*. **2004**, 57, 137–154.
28. Jauregi, E.; Lazkano, E.; Sierra, B. Approaches to Door Identification for Robot Navigation, Mobile Robots Navigation, Alejandra Barrera (Ed.), InTech, **2010**. ISBN: 978-953-307-076-6.
29. Niu, L.; et al. A Wearable Assistive Technology for the Visually Impaired with Door Knob Detection and Real-Time Feedback for Hand-to-Handle Manipulation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22-29 Oct. 2017; pp. 1500-1508.
30. Panchal, A.; Varde, S.; Panse, M. Character Detection and Recognition System for Visually Impaired People. In Proceedings of IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology, Bangalore, India, 20-21 May 2016; pp. 1492–1496.
31. Jabnoun, H.; Benzarti, F.; Amiri, H. Object Recognition for Blind People based on Features Extraction. In Proceedings of International Image Processing, Applications and Systems Conference, Sfax, Tunisia, 5-7 Nov. 2014; pp. 1-6.
32. Ciobanu, A.; Morar, A.; Moldoveanu, F.; Petrescu, L.; Ferche, O.; Moldoveanu, A. Real-Time Indoor Staircase Detection on Mobile Devices. In Proceedings of International Conference on Control Systems and Computer Science, Bucharest, Romania, 29-31 May 2017; pp. 287-293.
33. Nascimento, J.C.; Marques, J. S. Performance Evaluation of Object Detection Algorithms for Video Surveillance. *IEEE Transactions on Multimedia*. **2016**, 8(4), 761-774.
34. Hernandez, A.C.; Gómez, C.; Crespo, J.; Barber, R. Object Detection Applied to Indoor Environments for Mobile Robot Navigation. *Sensors* **2016**, 16, 1180.
35. Li, Z.; Dong, M.; Wen, S.; Hu, X.; Zhou, P.; Zeng, Z. CLU-CNNs: Object Detection for Medical Images. *Neurocomputing*. **2019**, 350, 53-59.
36. Baeg, S.; Park, J.; Koh, J.; Park, K.; Baeg, M. An Object Recognition System for a Smart Home Environment on the Basis of Color and Texture Descriptors. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, 29 Oct.-2 Nov. 2007; pp. 901-906.

37. Paletta, L.; Fritz, G.; Seifert, C.; Luley, P.; Almer, A. Visual Object Recognition for Mobile Tourist Information Systems. In Proceedings of the SPIE 5684, Multimedia on Mobile Devices, San Jose, California, USA, 14 March 2005; pp. 190-197.
38. Lu, S.; Wang, B.; Wang, H.; Chen, L.; Linjian, M.; Zhang, X. A real-time object detection algorithm for video. *Comput. Electr. Eng.* **2019**, *77*, 398–408, doi:10.1016/j.compeleceng.2019.05.009
39. Trabelsi, R.; Jabri, I.; Melgani, F.; Smach, F.; Conci, N.; Bouallegue, A. Indoor object recognition in RGBD images with complex-valued neural networks for visually-impaired people. *Neurocomputing* **2019**, *330*, 94–103, doi:10.1016/j.neucom.2018.11.032.
40. Malūkas, U.; Maskeliūnas, R.; Damaševičius, R.; Woźniak, M. Real Time Path Finding for Assisted Living Using Deep Learning. *J. Univers. Comput. Sci.* **2017**, *24*, 475–486, doi:10.3217/jucs-024-04-0475.
41. Jayakanth, K. Comparative Analysis of Texture Features and Deep Learning Method for Real-time Indoor Object Recognition. In Proceedings of the 2019 International Conference on Communication and Electronics Systems (ICCES); IEEE: Coimbatore, India, 2019; pp. 1676–1682.
42. Jabnoun, H.; Benzarti, F.; Morain-Nicolier, F.; Amiri, H. Video-based assistive aid for blind people using object recognition in dissimilar frames. *Int. J. Adv. Intell. Paradig.* **2019**, *14*, 122–139, doi:10.1504/IJAIP.2019.102967.
43. Kulikajevas, A.; Maskeliūnas, R.; Damaševičius, R.; Ho, E.S.L. 3D Object Reconstruction from Imperfect Depth Data Using Extended YOLOv3 Network. *Sensors* **2020**, *20*, 2025, doi:10.3390/s20072025.
44. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *ArXiv* **2016**, 1–10.
45. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *ArXiv* **2016**.
46. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor Segmentation and Support Inference from RGBD Images. In Proceedings of the Computer Vision – ECCV 2012; Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., Eds.; Springer: Berlin, Heidelberg, 2012; pp. 746–760.

47. Lai, K.; Bo, L.; Ren, X.; Fox, D. A large-scale hierarchical multi-view RGB-D object dataset. In Proceedings of the 2011 IEEE Int. Conference on Robotics and Automation; Shanghai, China, 2011; pp. 1817–1824.
48. ImageNet, Available online: http://www.image-net.org/ (accessed on Sep 4, 2020).
49. MCIndoor20000: A fully-labeled image dataset to advance indoor objects detection - Available online: https://www.sciencedirect.com/science/article/pii/S2352340917307424 (accessed on Sep 4, 2020).
50. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, 1–6.
51. Velazquez, R.; Varona, J.; Rodrigo, P.; Haro, E.; Acevedo, M. Design and Evaluation of an Eye Disease Simulator. *IEEE Latin America Transactions*. **2015**, 13(8), 2734-2741.
52. https://developer.nvidia.com/embedded-computing(accessed August 2020)
53. https://www.pyimagesearch.com/2020/03/25/how-to-configure-your-nvidia-jetson-nano-for-computer-vision-and-deep-learning(accessed August 2020)
54. NVIDIA Co. Jetson Partner Supported Cameras, 2020, Available at: https://developer.nvidia.com/embedded/jetson-partner-supported-cameras (accessed August 2020)
55. NVIDIA Co. Jetson Partner Hardware Products, 2020, Available at: https://developer.nvidia.com/embedded/community/jetson-partner-products (accessed August 2020)
56. NVIDIA Jetson Linux Developer Guide : Clock Frequency and Power Management, Available online: https://docs.nvidia.com/jetson/l4t/index.html#page/Tegra%2520Linux%2520Driver%2520Package%2520Development%2520Guide%2Fclock_power_setup.html%23 (accessed on Aug 31, 2020).
57. Özdemir, A.T. An Analysis on Sensor Locations of the Human Body for Wearable Fall Detection Devices: Principles and Practice. *Sensors* **2016**, *16*, 1–25, doi:10.3390/s16081161.
58. https://blogs.nvidia.com/blog/2016/10/27/wearable-device-for-blind-visually-impaired/(accessed August 2020)

59. Lin, T.Y.; et al. Microsoft COCO: Common Objects in Context. In Computer Vision – ECCV 2014; Fleet, D., Pajdla T., Schiele B., Tuytelaars T., Eds.; LNCS, 8693; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740-755. ISBN: 978-3-319-10601-4.
60. https://developer.nvidia.com/embedded/learn/get-started-jetson-nano-devkit (accessed August 2020)
61. https://docs.nvidia.com/deeplearning/frameworks/install-tf-jetson-platform/index.htm (accessed August 2020)
62. Alvarez-Pato, V.M.; Sanchez, C.N.; Dominguez-Soberanes, J.; Mendoza-Perez, D.E.; Velazquez, R. A Multisensor Data Fusion Approach for Predicting Consumer Acceptance of Food Products. *Foods* **2020**, 9, 774.
63. TensorFlow 1 Detection Model Zoo. Collection of pre-trained detection models. Available at: https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md (accessed August 2020)
64. Pissaloux, E.; Maybank, S.; Velazquez, R. On Image Matching and Feature Tracking for Embedded Systems: A State of the Art. In: *Advances in Heuristic Signal Processing and Applications*; Chatterjee, A., Nobahari, H.; Siarry, P, Eds.; Springer: Berlin, Germany, 2013, pp. 357-380. ISBN: 978-3-642-37879-9.
65. Visconti, P.; de Fazio, R.; Costantini, P.; Miccoli, S.; Cafagna, D. Innovative complete solution for health safety of children unintentionally forgotten in a car: a smart Arduino-based system with user app for remote control. *IET Science, Measurement & Technology*. **2020**, 14(6), pp. 665 – 675.
66. Visconti, P.; de Fazio, R.; Costantini, P.; Miccoli, S.; Cafagna, D. Arduino-based solution for in-car-abandoned infants' controlling remotely managed by smartphone application. *Journal of Communications Software and Systems*, **2019**, 15(2), pp. 89-100.
67. Liu W.; Anguelov D.; Erhan D.; Szegedy C.; Reed S.; Fu C.Y.; Berg, A. SSD: Single Shot MultiBox Detector. In: Leibe B.; Matas J.; Sebe N.; Welling M., Eds; LNCS 9905, Springer: Berlin, Germany, 2016, pp 21-37.
68. Redmon J.; Divvala S.; Girshick R.; Farhadi A. You only look once: unified, real-time object detection. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27-30 June 2016, pp. 779-788.

69. Xue Y.; Li Y. A fast detection method via region-based fully convolutional neural networks for shield tunnel lining defects. *Computer-Aided Civil and Infrastructure Engineering* **2018**, 33(8), 638-654.
70. Wang H.; Yu Y.; Cai Y.; Chen X.; Chen L.; Liu Q. A comparative study of state-of-the-art deep learning algorithms for vehicle detection. *IEEE Intelligent Transportation Systems Magazine* **2019**, 11(2), 82-95.
71. Intellica Co. A Comparative Study of Custom Object Detection Algorithms 2019, Available at: https://medium.com/@Intellica.AI/a-comparative-study-of-custom-object-detection-algorithms-9e7ddf6e765e (accessed August 2020)