Sistema de detección de objetos para personas con discapacidad visual

Diego Godoy Rojas - dfgodoyr@unal.edu.co — Luis Carlos Diaz - lcdiazf@unal.edu.co

RESUMEN

El presente proyecto propone el diseño de un algoritmo funcional para el sistema de detección de objetos asistido para personas con discapacidades visual que facilitara su día a día de manera segura y confiable. El software permitirá reconocer objetos en un rango de visión frontal y de esta manera poderlos comunicar al usuario con limitaciones visuales mediante respuestas auditivas. Para cumplir con las metas planteadas se empleó la metodología estructurada de diseño y desarrollo de productos, en la que se abarcan diferentes etapas y actividades que orientan el proyecto desde su conceptualización hasta su puesta en marcha y funcionamiento. El resultado final del proyecto dentro del marco del curso se estableció hasta el desarrollo del componente referido a visión por ordenador (Procesamiento, segmentación, detección y clasificación), dejando el componente de respuesta auditiva como un alcance a futuro que se tratara de implementar.

I. INTRODUCCIÓN

La discapacidad visual se refiere a una condición dada por la pérdida de visión que constituye una limitación significativa de la capacidad visual, esta se da como resultado de una enfermedad, un traumatismo o una condición congénita o degenerativa que no puede ser corregida por medios convencionales. Según estimaciones recientes de la Organización Mundial de la Salud (OMS), 285 millones de personas en todo el mundo tienen problemas de visión. De ellas, 39 millones son ciegas y 246 millones tienen baja visión. En Colombia aproximadamente hay registradas 1.948.332 personas con discapacidad visual. La pérdida de visión en cualquier grado tiene un efecto debilitante en la capacidad de un individuo para realizar actividades cotidianas como la navegación, la evasión de obstáculos, el acceso a la información, el análisis y reconocimiento de las señales de comunicación no verbal (como el contacto visual, las expresiones faciales, los gestos con las manos, la postura corporal, entre otros.) y el reconocimiento de los objetos animados e inanimados del entorno. Además, la pérdida de visión aumenta el riesgo de sufrir lesiones no intencionadas, como caídas, lesiones relacionadas con el trabajo o con el tráfico tanto vehicular como peatonal.

Aunque las personas con discapacidad visual, debido a la naturaleza de esta, tienden a ser muy organizados y a guardar los objetos importantes, como carteras y llaves, en lugares muy concretos de sus casas y lugares de trabajo, no se puede hacer lo mismo con todos los objetos. Además, los objetos pueden extraviarse o cambiar inesperadamente de posición (por ejemplo, al caerse accidentalmente o ser movidos por otra persona) de esta manera la persona puede pasar varias horas al día buscando objetos como la grapadora de su escritorio, un lápiz que puede haberse caído al suelo y así con muchos objetos más. Este problema se agrava cuando entra en un lugar desconocido, como un restaurante, un banco o una tienda, donde no conoce la ubicación de los distintos objetos, por lo que su memoria ya no puede ayudarle y sus sentidos restantes sólo pueden proporcionarle información limitada sobre los pocos objetos del entorno que son directamente accesibles a esos sentidos. Pues con un sentido de la vista muy debilitado o inexistente, una persona con discapacidad visual tiene que depender de su memoria y confiar mucho en sus otros sentidos, es decir, el oído, el tacto, el gusto y el olfato, para localizar e identificar los objetos de su entorno. Dado que la cantidad de información que se recibe a través del sistema visual humano es mucho mayor y la velocidad a la que se procesa es considerablemente más rápida en comparación con la información que se recibe a través de los otros sentidos, la dependencia de estos medios alternativos de percepción hace que se dedique una cantidad de tiempo significativamente mayor a esta tarea. Además, en ciertos casos, el individuo puede querer distinguir entre varios objetos que son similares en tamaño, forma y textura. Si la única diferencia entre estos objetos es su codificación visual, la retroalimentación auditiva, táctil u olfativa no ayudaría a discriminarlos.

Lo mencionado anteriormente pone en manifiesto la necesidad de desarrollar dispositivos que ayuden a las personas que sufren de este tipo de discapacidades a mejorar su movilidad y su capacidad de interacción con el entorno, al tiempo que les permitan ser más autosuficientes.

La visión de máquina es una alternativa importante en el desarrollo de estos dispositivos, se han dedicado bastantes esfuerzos en la detección de objetos usando diferentes técnicas que han cambiado a través del tiempo, una primera alternativa es el reconocimiento de patrones obtenidos del procesamiento de imágenes, estos patrones pueden ser espaciales o frecuenciales, en este caso el reconocimiento se da cuando se compara con modelos obtenidos previamente, usando técnicas de probabilidad. Esta alternativa muestra buenos resultados en cuanto precisión, sin embargo, tiene un costo computacional elevado a causa de todos los posibles puntos de vista que puede tener un objeto, además requiere una representación tridimensional, que usualmente no se tiene. Adicionalmente las

alternativas de este tipo suelen ser bastante sensibles al ruido, perturbaciones o interferencias entre los diferentes objetos.

Por otro lado, la inteligencia artificial brinda soluciones con un mejor rendimiento, y que pueden facilitar el proceso, además de ser más robusta ante situaciones inesperadas. En este caso se requiere un entrenamiento de la red neuronal, usando imágenes de prueba que han sido procesadas para obtener regiones y características de interés, para que el sistema sea capaz de distinguir objetos en una imagen. En esta fase en entrenamiento se requieren imágenes que estén previamente clasificadas y anotadas por lo que se sugiere el uso de una base de datos.

Teniendo en cuenta estos antecedentes surge el marco de este proyecto, en donde se busca implementar técnicas de detección de objetos, de tal forma que se obtenga un prototipo final capaz de trabajar en entornos no controlados, donde existen interferencias, superposiciones entre objetos y desorden, además, se articulara este sistema de detección para realizar una conversión de los resultados obtenidos a voz.

De esta manera en el presente artículo haremos una exploración mediante un estado del arte de los distintos acercamientos que ha tenido este campo de trabajo, para posteriormente estructurar la metodología de trabajo a seguir durante el marco de la materia para el adecuado cumplimiento de este. Seguido a esto se explicará el desarrollo e implementación técnica del proyecto y de esta manera poder llegar a la obtención de resultados para su análisis y posterior conclusión.

II. ESTADO DEL ARTE

Al problema reconocimiento de objetos se han dado distintas aproximaciones pensando en aplicaciones específicas como reconocimiento de entorno para movimiento autónomo de vehículos, verificación de calidad en procesos industriales y también en la creación de sistemas para personas invidentes. También hay una gran cantidad de trabajos en la detección de objetos genéricos basados en la comparación con modelos adquiridos previamente usando distintas técnicas.

La mayoría de estas técnicas tiene dos fases, una de aprendizaje o adquisición de modelos, donde se recogen imágenes que luego de alguna transformación o se almacenan de tal forma que puedan ser usadas en la fase de detección. En esta fase a grandes rasgos lo que se busca aislar e identificar cada objeto, esto se puede lograr con distintos métodos que revisan distintas características de la imagen y la compara con los datos almacenados para determinar la coincidencia.

En el caso de [1] se usan tres características de la imagen, profundidad de píxel, normales a la superficie y curvatura de la superficie. Luego se obtienen los histogramas de estas características y se comparan con los modelos mediante métodos probabilístico. El algoritmo propuesto usa características locales de la imagen y sin necesidad de segmentación, esta propuesta fue evaluada usando imágenes tridimensionales sintéticas con oclusión parcial causada por la sombra del mismo objeto. Este sistema mostró una precisión del 93 % con imágenes ideales y del 89 % con imágenes con una oclusión del 20 %. Una de la ventaja de este sistema es que ahorra el proceso de segmentación, pero no es efectivo ante oclusiones causadas por otros objetos o en presencia de fondos desordenados, por esta razón no es viable en este proyecto, pero es importante tener en cuenta las características de las imágenes usadas por el algoritmo para posteriores pruebas.

A diferencia de la propuesta anterior en [2] se muestra una opción que no usa la profundidad ya que a veces no se cuenta con esta medida, el hardware para realizar medidas de profundidad puede ser inviable o si se usan métodos para determinar la profundidad basándose en el brillo de cada píxel se puede presentar problemas con el fondo o la presencia de otros objetos como en el ejemplo anterior. En este caso se busca determinar correlaciones que agrupan bordes y que sean invariantes al punto de vista, con estos grupos se seleccionan posibles candidatos entre los modelos tridimensionales previamente almacenados desde diferentes puntos de vista hasta encontrar uno que sea acertado. De este proceso es importante resaltar el proceso de segmentación de bordes en líneas rectas unidas por puntos y las correlaciones espaciales, colinealidad, paralelismo y proximidad de puntos finales.

Otra opción usando modelos en 3D se presenta en [3], donde se propone un algoritmo para la creación de los modelos a partir de imágenes del mismo objeto desde muchos puntos de vista, además se propone un algoritmo que recibe una imagen donde hay varios objetos, incluso superpuestos y es capaz de identificarlos. Esto se logra mediante una búsqueda exhaustiva de correspondencias de bordes y superficies entre la entrada y los modelos, de esta manera se eligen las opciones con más correspondencias y se repite el proceso hasta elegir una opción que se compara con la imagen para evitar falsos positivos, si esta comprobación es positiva se procede a separar el objeto de imagen original. Esta opción es muy robusta y contiene información interesante sobre obtención de los modelos tridimensionales, algoritmos para encontrar coincidencias y verificar resultados. Sin embargo, tiene un costo computacional elevado debido a que la búsqueda exhaustiva requiere muchas comparaciones por cada objeto, además se requiere que los objetos a identificar estén plenamente descritos por el modelo.

Una solución más cercana a la intención del proyecto se presenta en [4], donde se usa una máquina de vectores de soporte (SVM) para encontrar coincidencias en las imágenes, realizaron pruebas en imágenes reales que contienen rostros humanos, carros, motocicletas, aeroplanos y hojas de árboles y los resultados muestran una precisión del setenta al noventa y cuatro por ciento. Tanto la fase de aprendizaje como de detección hay tres pasos importantes, en primer lugar, se identifican puntos de interés usando un detector Kadir-Brady basado en derivadas (DKB), después se extraen características de las regiones cercanas a los puntos de interés usando la transformada curvelet y filtro de Gabor, para luego ser enviadas al SVM que después del entrenamiento es capaz de distinguir los objetos presentes.

En [5] se propone una integración de detección de objetos a un sistema de narración visual para ayudar a personas invidentes, esta implementación puede ser usada en dispositivos comunes como computadores, teléfonos celulares y tablets. Este sistema usa una implementación del sistema de detección YOLO que usa redes neuronales para la detección de objetos y una librería de voz para la narración del entorno. Esta implementación es bastante asequible y tiene una precisión de al menos 60 %.

El entrenamiento y pruebas de un sistema de detección de objetos requiere un conjunto de datos, para esto se propone Open Images Dataset V6 de Google [6], esta base de datos contiene aproximadamente nueve millones de imágenes con anotaciones de etiquetas, cuadros delimitadores, segmentación de objetos, relaciones visuales y narrativas localizadas. Las imágenes contienen 600 clases de objetos y son diversas en cuanto a complejidad, en promedio tiene 8.3 objetos por imagen.

III. METODOLOGÍA

El proceso de desarrollo del proyecto para lograr la asistencia a personas con limitaciones visuales de manera efectiva se compone de dos partes fundamentales, en primer lugar se debe realizar la adecuada detección de objetos, para esto se usó la biblioteca de código abierto TensorFlow (TF), una plataforma de aprendizaje de máquina flexible y fácil de usar que se puede emplear en diversas aplicaciones específicas. La segunda parte se remite al apoyo auditivo sobre el objeto que se está detectando, para esto se usó la librería de Python *Python text-to-speech x-platform* (pyttsx3), la cual puede ser usada de manera offline y es compatible con la versión de Python que usa la versión más reciente de TensorFlow.

TensorFlow es una plataforma que facilita el uso de redes neuronales mediante aprendizaje profundo, este puede ser usado en la detección de objetos mediante Transfer Learning o entrenamiento basado en un modelos pre-entrenados, estos modelos pre-entrenados ya se encuentran en las librerías de detección de objetos de TensorFlow, para el proyecto se empleó el modelo "***SSD MobileNet v2 320x320***", el cual tiene una velocidad de respuesta de 19ms, una de los más rápidas dentro de la librería de detección de objetos de TensorFlow.

El entrenamiento del modelo requiere de un conjunto de datos amplios formado por diversas imágenes de los objetos que se desean detectar, estas imágenes deben presentar variaciones en su tamaño, iluminación, posición, entre otras características, el proyecto se centra en la detección de cuatro objetos que para personas con limitaciones visuales se puede dificultar su identificación y diferenciación, como lo es el caso de tarjetas y billetes, de esta manera, el conjunto de datos del proyecto está formado por doscientas imágenes para cada una de las cuatro clases, dos billetes de diferente denominación (2000 y 5000), un documento de identidad y una tarjeta de crédito. Cada imagen del conjunto de datos es respectivamente etiquetada para el objeto que contiene y este se delimita dentro de un cuadro que aporta información sobre el objeto, para esta tarea se usó la implementación de interfaz gráfica en Python LabelImg disponible en [7], la cual brinda las herramientas necesarias para realizar los procesos descritos como se observa en la figura 1. Una vez se tiene listo el conjunto de datos, se divide en los conjuntos de prueba y entrenamiento, para ser comprimidos y subidos a la nube para entrenar el modelo en la nube de Google Colaboratory ya que permite usar un entorno de ejecución con gpu y de esta manera poder acelerar el proceso de entrenamiento.

Figura 1: Interfaz gráfica de LabelImg.

Para la construcción del modelo de aprendizaje profundo se utilizaron configuraciones estándar de redes neuronales, donde se especificaron las estructura de la red, el preprocesamiento que se realiza a las imágenes de entrada y el extractor de características, en este caso se usó *single shot detection mobilenet* (ssd mobilenet v2 fpn keras). Una vez realizada la configuración mencionada se lleva a cabo el entrenamiento, el cual da como resultado en un modelo portable y listo para reconocer los objetos pertenecientes a las clases definidas anteriormente. Todo el proceso para llegar a este punto se puede ver ilustrado a continuación:

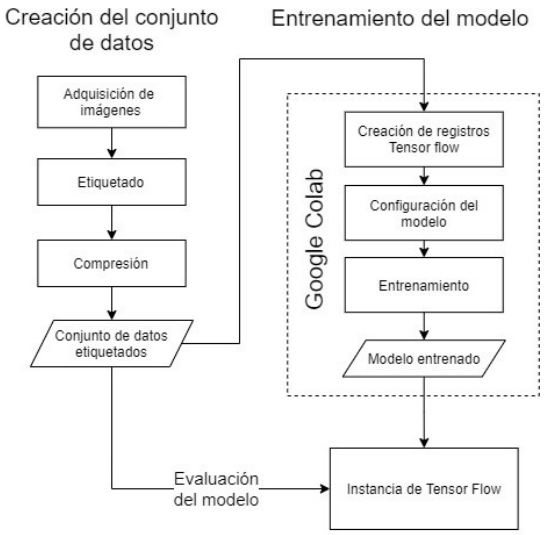

Figura 2: Diagrama de flujo del entrenamiento.

Una vez el sistema de detección de objetos está listo se procede a la implementación de la adquisición de imágenes en tiempo real usando la función VideoCapture de la librería OpenCV, con la cual en cada iteración del sistema se obtiene una imagen de la cámara empleada, que se entrega a una instancia de TF configurada con el modelo obtenido anteriormente. El modelo retorna el número de identificación (ID), una caja delimitadora (bbox) y un puntaje de cada objeto detectado en la imagen.

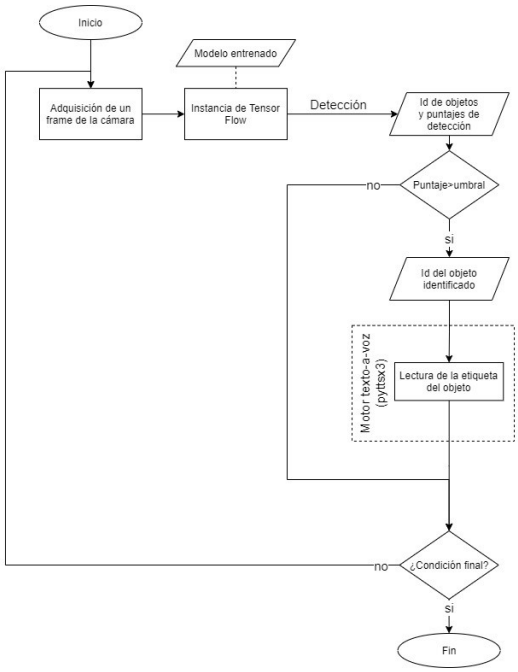

Figura 3: Diagrama de flujo del sistema completo.

Finalizada la detección se obtiene la etiqueta correspondiente del objeto con el mejor puntaje de detección, la cual es interpretada y leída por el sistema de texto a voz de pyttsx3, el cual tiene un módulo para español. En el último paso de un ciclo del programa se verifica si hay una condición para cerrar el sistema, sino se retorna al inicio donde se obtiene un nuevo frame con el cual se repiten todos los pasos. Cada paso del algoritmo se muestra en el diagrama de flujo de la figura 3.

IV. RESULTADOS

Una vez cargado y compilado el modelo, se procede a realizar pruebas de detección en imágenes, en la figura 4 podemos observar cómo se realiza la adecuada distinción entre las dos denominaciones de billetes detectando el billete de 2000 con una precisión del 90 % mientras que el billete de 5000 se detecta con una precisión menor de 85 %, en la figura 5 vemos como el modelo es capaz de detectar diversos objetos de una misma clase en una sola imagen, ambas con una precisión mayor o igual al 90 %.

Figura 4: Detección simultanea de dos billetes de diferente denominación.

Figura 5: Detección simultanea de dos billetes de igual denominación.

El otro par de objetos idénticos en forma corresponde a las tarjetas plásticas, vemos como en la figura 6 el modelo es capaz de detectar y distinguir cada uno de estos objetos dentro de una imagen, por último, en la figura 7 podemos comprobar la eficiencia del modelo para distinguir dos objetos totalmente diferentes en una misma imagen. A partir de esto se tuvo un conjunto de validación correspondiente a 100 imágenes de cada clase, donde se obtuvo una tasa de verdaderos positivos del 83 % aproximadamente, por lo tanto se aproxima la precisión del modelo diseñado a 83 %, debido a que uno de los objetivos del proyecto es su realización en tiempo real (Vídeo) con ayuda auditiva (Sonido), se imposibilita plasmar en el presente documento soportes sobre su funcionamiento, para esto se dejara anexado en el repositorio del proyecto tres vídeos que detallan los demás resultados obtenidos.

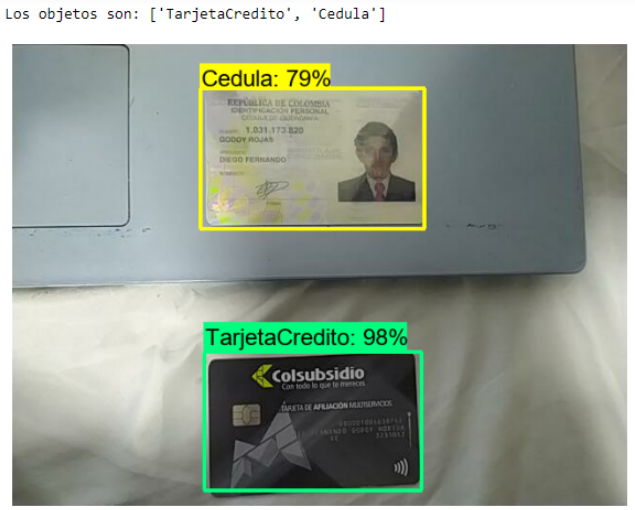

Figura 6: Detección simultanea de dos tarjetas plásticas de diferente uso.

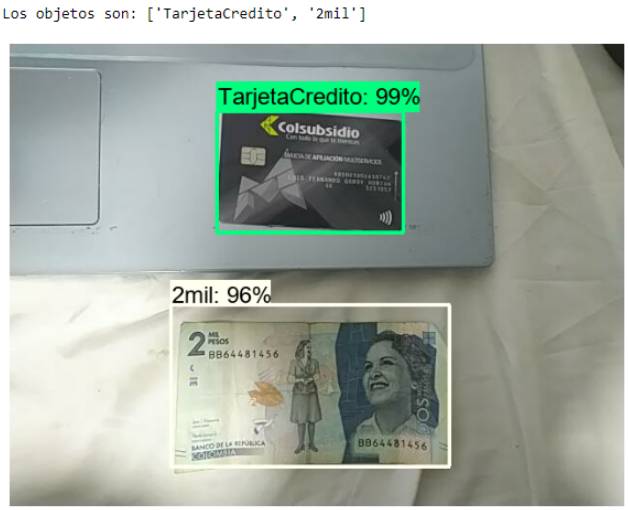

Figura 7: Detección simultanea de dos elementos de diferente clase.

V. CONCLUSIÓN

El sistema de asistencia de reconocimiento de objetos para personas con limitaciones visuales es un gran apoyo para esta población al brindarle un modo de poder distinguir entre objetos físicamente similares de manera rápida y confiable, este sistema presenta una gran diferencia respecto a los trabajos descritos en el estado del arte debido a su versatilidad ya que permite un fácil entrenamiento del modelo a partir de los objetos y entornos cotidianos de la persona que lo vaya a emplear sin requerir un gran número de datos para su entrenamiento.

Debido a la limitaciones establecidas a lo largo del documento, la optimización del proyecto recae en la solución de las mismas, pues el proyecto tal cual como se diseñó, puede seguir evolucionando y mejorando en trabajos futuros que permitan su escalabilidad a otras plataformas como microprocesadores, microcontroladores, dispositivos Android, entre otros, haciendo uso del Framework de construcción de Tensorflow para dispositivos TFLite, lo cual permitiría una mayor comodidad en su uso y la posible implementación en algún tipo de montura para gafas.

REFERENCIAS

[1] G. Hetzel, B. Leibe, P. Levi, B. Schiele, "3D Object Recognition from Range Images using Local Feature Histograms", IPVR University of Stuttgart & Perceptual Computing and Computer Vision Group ETH Zurich, 2001.

[2] D. Lowe, "Three-Dimensional Object Recognition from Single Two-Dimensional Images", New York University, New York, 1987.

[3] A. Mian, M. Bennamoun, R. Owens, "Three-Dimensional Model-Based Object Recognition and Segmentation in Cluttered Scenes", IEEE Transactions on pattern analysis and machine intelligence, vol. 28, no. 10, pp 1584-1601 ,october 2006

[4] Bhuvaneswari, R., Subban, R., Novel Object Detection and Recognition System based on Points of Interest Selection and SVM Classification, *Cognitive Systems Research* (2018), doi: https://doi.org/10.1016/j.cogsys.2018.09.022

[5] J. Nasreen, W. Arif, A. Shaikh, Y. Muhammad, "Object Detection and Narrator for Visually Impaired People" in 6th IEEE International Conference on Engineering Technologies and Applied Sciences, 2019.

[6] Google, Open Images Dataset V6 + Extensions [Online], Available: https://storage.googleapis.com/openimages/web/index.html

[7] tzutalin, Label Img [Online], Available: https://github.com/tzutalin/labelImg