Sistema basado en reconocimiento de objetos para el apoyo a personas con discapacidad visual (¿Que tengo enfrente?)

System based on object recognition to support people with visual disabilities

Carlos Aramendiz U*, David Escorcia G*, Jesus Romero C*, Kevyn Torres R*, Carlos Triana P*, Silvia Moreno-Trillos**

{smoreno12@unisimonbolivar.edu.co}

Universidad Simón Bolívar, Barranquilla-Colombia

Resumen

Se sabe que las personas que sufren de una limitante visual o en su defecto son discapacitados visuales tienden a enfrentarse a retos como no saber el objeto ante ellos o reconocer el entorno en que se encuentran; por tal motivo estas personas suelen tener casi siempre un acompañante que los guíe en su mundo de oscuridad. Pero, en el caso de que esta persona no se encuentre, ya sea porque la persona discapacitada no cuenta con los recursos para contratar a alguien, o no tenga personas cercanas disponibles cuando él/ella lo necesite, es necesario alguna ayuda tecnológica que supla esa falencia.

En este proyecto se propone el desarrollo de una aplicación móvil que sirva de apoyo a las personas con discapacidad visual. Esta aplicación tendrá la capacidad de reconocer objetos a partir de una cámara móvil y brindará esta información a la persona de manera audible.

Palabras clave:

Motor Visual, Redes Neuronales, Redes Neuronales por Convolución, Visión Computacional, Discapacidad visual, Aplicaciones Móviles, Reconocimiento de Objetos.

Abstract

It is known that people with visual impairment tend to face challenges such as not knowing the object before them or recognizing the environment in which they find themselves; For this reason these people usually have a companion to guide them in their world of darkness. But, in the event that this person is not around, either because the disabled person does not have the resources to hire someone, or does not have close people available when he/she needs them, some technological help is needed to overcome this shortcoming.

This project proposes the development of a mobile application to support the visually impaired. This application will have the ability to recognize objects from a photo and will provide this information to the person audibly.

Key-words:

Neural Networks, Convolution Neural Networks, Computational Vision, Visual Disability, Mobile Applications, Object Recognition.

Para referenciar este artículo (IEEE):

[N]C. Aramendiz*, D. Escorcia*, J. Romero*, K. Torres, C. Triana & S. Moreno-Trillos, “Sistema basado en reconocimiento de objetos para el apoyo a personas con discapacidad visual”, *Investigación y Desarrollo en TIC,* vol. 7, no. 2, pp. 55-60 201

I. INTRODUCCIÓN

Uno de los principales problemas que tiene la tecnología en la sociedad es que no está al alcance de todo tipo de personas dejando de lado temas como la pobreza y la localidad, están las discapacidades físicas en el caso más específico; aquellas personas que tienen dificultad en la vista por problemas de salud o en su defecto aquellas que quedaron totalmente privadas del sentido de la vista.

El objetivo de este estudio es determinar cómo hacer más fácil el día a día de las personas con discapacidad visual y así contribuir de una manera indirecta a mejorar su calidad de vida y que se mantengan informados acerca del entorno en el que se encuentran. Con este fin, la pregunta de investigación es la siguiente: ¿Cómo podemos contribuir de una manera tecnológica a facilitar el día a día de una persona con discapacidad visual?

Con el fin de ayudar a las personas y de responder a la pregunta de investigación anteriormente dada, nos damos a la tarea de diseñar, crear y gestionar un software (aplicativo móvil) cuya función sea darle ayuda o soporte a las personas invidentes.

Una de las tareas más complejas a la hora de abordar proyectos que integren técnicas avanzadas de IA es el desarrollo de los modelos y la puesta en producción de los mismos. En el caso de las técnicas de reconocimiento de imágenes, una alternativa muy interesante es apoyarnos en los modelos pre-entrenados que las principales plataformas cloud nos ofrecen; servidores como image.net al igual que las diferentes librerías y archivos que están alojados en internet que nos sirven de cierta manera a agilizar los procesos de reconocimiento.

Sabemos que el ser humano ha tratado de emular lo que vendría siendo teóricamente el ojo humano, pero más concretamente su visión, creando cámaras cada vez con mejor definición y más potentes; así que con el uso de estas cámaras (de celular) se puede “escanear” o percibir el objeto que se requiere procesar, luego nuestra aplicación móvil emite una serie de imágenes simultáneas captadas por la cámara haciendo un reconocimiento de imagen por medio de la de redes neuronales concretamente con YOLO 3, mostrando el resultado en tiempo real a la persona con discapacidad visual por voz.

Usaremos redes neuronales ya pre-entrenadas, mencionadas por su efectividad y agilidad.

II. MARCO TEÓRICO

Visión Computacional

El reconocimiento de objetos es una técnica que se utiliza para reconocer efectivamente un objeto en la imagen tratando de imitar en las computadora la capacidad que tienen nuestro ojos. Es decir, trata de interpretar las imágenes recibidas por dispositivos como la cámara y

reconocer los objetos, ambiente y posición en el espacio.[1]

Su función principal es reconocer y localizar objetos en el ambiente mediante el procesamiento de las imágenes. La visión computacional estudia estos procesos, para entenderlos y así elaborar máquinas con capacidades similares. Existen varias definiciones de visión, entre estas podemos mencionar las siguientes:

- "Visión es saber que hay y donde mediante la vista" (Aristóteles)
- "Visión es recuperar de la información de los sentidos (vista) propiedades válidas del mundo exterior" (Gibson)
- "Visión es un proceso que produce a partir de las imágenes del mundo exterior una descripción que es útil para el observador y que no tiene información irrelevante". (Marr)

Las tres son esencialmente válidas, pero la que tal vez se acerca más a la idea actual sobre visión computacional es la definición de Marr. En esta definición hay tres puntos importantes que hay que tener presentes: (i) visión es un proceso computacional, (ii) la descripción a obtener depende del observador y (iii) es necesario eliminar la información que no sea útil (reducción de información).

En la actualidad existen múltiples aplicaciones prácticas de la visión computacional, entre estas podemos mencionar las siguientes:

- Robótica móvil y Vehículos autónomos. Se utilizan cámaras y otros tipos de sensores para localizar obstáculos, identificar objetos y personas, encontrar el camino, etc.
- Manufactura. Se aplica visión para la localización e identificación de piezas, para control de calidad, entre otras tareas.
- Interpretación de imágenes aéreas y de satélite. Se usa el procesamiento de imágenes y visión para obtener mejores resultados de las imágenes obtenidas, para identificar diferentes tipos de cultivos, para ayudar en la predicción del clima, etc.
- Análisis e interpretación de imágenes médicas. La visión se aplica para ayudar en la interpretación de diferentes clases de imágenes médicas como rayos-X, tomografia, ultrasonido, resonancia magnética y endoscopia.
- Interpretación de escritura, dibujos, planos. Se utilizan técnicas de visión para el reconocimiento de textos, lo que se conoce como reconocimiento de caracteres. También se aplica a la interpretación automática de dibujos y mapas.
- Análisis de imágenes microscópicas. El procesamiento de imágenes y visión se utilizan para ayudar a interpretar imágenes microscópicas en química , física y biología .
- Análisis de imágenes para astronomía. Se usa la visión para procesar imágenes obtenidas por telescopios, ayudando a la localización e identificación de objetos en el espacio.
- Análisis de imágenes para comprensión Aunque la compresión de imágenes ha sido normalmente una sub área del procesamiento de imágenes, recientemente se están desarrollado técnicas más sofisticadas de compresión que se basan en la interpretación de las imágenes.[1]

Redes neuronales convolucionales (CNN)

El aprendizaje profundo o Deep Learning (DL) permite que modelos computacionales compuestos por varias capas de procesamiento puedan aprender representaciones sobre datos con múltiples niveles de abstracción y, mediante ese concepto, hallar representaciones precisas de forma autónoma en grandes volúmenes de datos. El DL ha logrado actualmente grandes avances en el reconocimiento de imágenes y video. Un caso particular de DL son las redes convolucionales o Convolutional Neural Networks (CNN), las cuales definen actualmente el estado del arte de varios problemas de visión computacional, dado su buen desempeño y problemas de reconocimiento e interpretación en imágenes y video[2]. Su capacidad para actuar adecuadamente en estos contextos está basada en características fundamentales: conexiones locales, pesos compartidos, pooling y el uso de una gran cantidad de capas[3]. El propósito de CNN es poder sacar todas las características de una imagen y luego usar estas características obtenidas para detectar o clasificar los objetos en una imagen. Los parámetros de los filtros que se pueden aprender en estas capas; se ajustarán y optimizarán junto con los componentes de clasificación para reducir el error de clasificación total. La aparición de CNN ha llevado a un rápido desarrollo del campo de detección de objetos. Por esto nos permitimos mencionar los algoritmos avanzados y más recientes en el campo de detección de objetos con redes neuronales convolucionales: Faster R-CNN, Single Shot Multibox Detector (SSD)[4] y You Only Look Once (YOLO), que es la adoptada en este trabajo y se describe en detalle a continuación [2].

LA RED YOU ONLY LOOK ONCE (YOLO)

En YOLO se toma la detección de objetos como un problema único de regresión, una única red convolucional predice simultáneamente múltiples cuadros delimitadores que enmarcan los objetos en la imagen y predice probabilidades condicionales por cada clase p(Clase | Objeto) para cada uno de estos cuadros delimitadores. La red neuronal puede lograr una velocidad de ejecución de 45 fotogramas por segundo (fps) en computadoras de propósito general. [5]

Figura 2: Proceso de YOLO, de izquierda a derecha: (1) división en bloques 13x13, (2) realización de predicciones, y (3) umbralización para obtener solo las detecciones más fehacientes [2].

YOLO trabaja globalmente sobre la imagen cuando hace predicciones, a diferencia de la técnica de ventana deslizante y las técnicas basadas en el análisis de las regiones en una imagen. Por esto, codifica implícitamente la información contextual, modela el tamaño y la forma de los objetos, así como su apariencia [2] (véase Fig. 2).

Como se puede apreciar en la Fig. 3, YOLO en su primera versión tiene 24 capas convolucionales seguidas por 2 capas completamente conectadas, y en lugar de los módulos iniciales propuestos por GoogleNet, YOLO utiliza capas de reducción de 1×1 seguidas de capas convolucionales de 3×3. En

la siguiente versión YOLO9000, al agregar la normalización de lotes en todas las capas convolucionales, se obtiene una mejora de más del 2% en la precisión promedio (mAP, por sus siglas en inglés). La normalización por lote también ayuda a regularizar el modelo y se elimina el sobreajuste. YOLO9000 predice las detecciones en un mapa de características de 13×13. Si bien esto es suficiente para objetos grandes, debe beneficiarse de funciones de grano más fino para localizar objetos más pequeños. [6]

III. ESTADO DEL ARTE

Existe un trabajo de Detección de Equipos de Protección Personal Mediante Red Neuronal Convolucional, en el que se aplicaron Redes CNN (REDES NEURONALES CONVOLUCIONALES), YOLO (YOU ONLY LOOK ONCE). Desarrollando un dataset de trabajadores Metalúrgicos. "*Para el proceso de entrenamiento, se utilizó la distribución de YOLO9000 original, a través de Darknet [6]. Para el proceso de entrenamiento se utilizó un ordenador con 16 GB de memoria, una tarjeta gráfica NVIDIA GeForce GTX 1080 TI y un procesador Intel Core i5. El proceso de entrenamiento fue relativamente sencillo, ya que, está muy bien explicado en la página de los autores.*"[7]

Se introdujo YOLO9000, un sistema de última generación, en tiempo real sistema de detección de objetos que puede detectar más de 9000 objetos categorías.[8]

El método de detección de YOLO, tanto el novedoso como el extraído de anteriores trabajos. El modelo mejorado, YOLOv2, es lo último en tareas de detección estándar como PASCAL VOC y COCO. Usando un novedoso método de entrenamiento a múltiples escalas, el mismo YOLOv2.[8]

El modelo puede funcionar en varios tamaños, ofreciendo un fácil intercambio entre la velocidad y la precisión. A 67 FPS, YOLOv2 consigue 76,8 mAP en VOC 2007. A 40 FPS, YOLOv2 obtiene 78,6 mAP, superando los métodos de última generación como RCNN más rápido con ResNet y SSD mientras sigue funcionando significativamente más rápido. Por último, se propone un método para entrenar conjuntamente en la detección y clasificación de objetos. Usando este método se entrenó YOLO9000 simultáneamente en el conjunto de datos de detección de COCO y el conjunto de datos de clasificación de ImageNet. El entrenamiento conjunto permite a YOLO9000 predecir las detecciones para las clases de objetos que no tienen datos de detección etiquetados. Se valida la tarea de detección de ImageNet. YOLO9000 consigue 19,7 mAP en el conjunto de validación de la detección de ImageNet a pesar de que sólo tiene datos de detección para 44 de las 200 clases. En las 156 clases que no están en el COCO, YOLO9000 obtiene 16.0 mAP. YOLO9000 predice detecciones para más de 9000 diferentes categorías de objetos, todo en tiempo real.[8]

Otras de las implementaciones que se llevaron a cabo fueron la de estudios de Redes Neuronales Convolucionales para el Reconocimiento Automático de Imágenes de macroinvertebrados. Se realizó una evaluación preliminar en la tarea de clasificación utilizando el modelo Inception-v3. Para la construcción del modelo se utilizaron las librerías Tensorflow y Keras, las cuales permiten aplicar funciones matemáticas para el procesado de los datos y el entrenamiento del modelo. Para analizar variantes en el modelo utilizamos TensorBoard,

un complemento que permite graficar los elementos que conforman el sistema y estudiar el comportamiento de los modelos utilizados.[9]

Existen varios trabajos previos que aplican detección de objetos. En el artículo de Ansari "Review of Deep Learning Techniques for Object Detection and Classification" mencionan varios enfoques. Krizhevsky et al. [10] propusieron la técnica para la clasificación de objetos. Realizan tarea de clasificación en 1,28 millones de imágenes que pertenecen a 1000 clases. En esta técnica, utiliza CNN para la clasificación de objetos. Utilizan 5 capas convolucionales y 3 fully-connected capas. Utilizan diferentes tamaños de filtro en diferentes capas convolucionales con diferente stride. Alexnet obtiene 57.0% de precisión para el top-1, mientras que para el top-5 obtiene 80,3% de precisión. Simonyan y Zisserman [11] realizan tareas de clasificación en 1,3 millones imágenes que pertenecen a 1000 clases. En esta técnica, utiliza CNN para el objeto clasificación. Hacen la red que contiene 19 capas de las cuales 16 son el la capa convolucional y 3 son la capa fully-connected. Utilizan el tamaño muy pequeño del filtro a toda la capa convolucional con un stride. VGG obtiene 70,5% de precisión para la parte superior-1 mientras para el top-5 obtiene una precisión del 90,0%.

IV. MATERIALES Y MÉTODOS

En el Sistema basado en reconocimiento de objetos para el apoyo a personas con discapacidad, utilizamos las librerías de OpenCV para Android y YOLO V3.

Una vez hecha la integración de OpenCV en el proyecto de Android "QueTengoEnFrente.apk", se crearon los diferentes componentes de la Vista:

-<org.opencv.android.JavaCameraView/>

-<Button/>

El sistema comienza en acción cuando se presiona el botón de "ON".

El sistema de reconocimiento, al trabajar con la Red Convolucional YOLO en su V3 es capaz de reconocer objetos a partir de una entrada de vídeo. En esta parte, entra en juego la librería de OpenCV la cual toma la red pre-entrenada y comienza a determinar los bordes de las figuras que se encuentran en escena (en tiempo real) y los pasa por las diferentes capas de esta.

Después de reconocer el objeto en escena, el nombre del mismo (tomado de una lista preconfigurada y traducida al español), es pasado a una "cola" de texto incluida en la Librería TTS (TextToSpeech) y de la cual se hizo una pequeña API para acceder más fácil a sus herramientas.

Por último se envía el resultado a una salida de Altavoz con la pronunciación del objeto reconocido.

*Nota importante: Los recursos como la red pre-entrenada, deben estar en la carpeta "dnns" ubicada en la raíz por ejemplo: (**storage/emulated/0/dnns/yolov3-tiny.cfg),** del celular a usar la APP. Esta carpeta puede ser descargada desde el siguiente link:

https://drive.google.com/file/d/1ZSrJVDsce3THDuaGqzTsHCfA_572KZ4g/view?usp=sharing

Si el link no está disponible puede escribir a:

compumarana@gmail.com

Versiones de la App:

https://drive.google.com/drive/folders/10w1b0vyb3O5CL475KHMn-JJakwdrwuzj?usp=sharing

V. RESULTADOS

El prototipo contaba con una pantalla inicial amigable al usuario donde solo hay un botón con el cual activa o desactiva el reconocimiento de objetos, pero el botón no era lo suficientemente grande para ser usada por una persona con discapacidad visual.

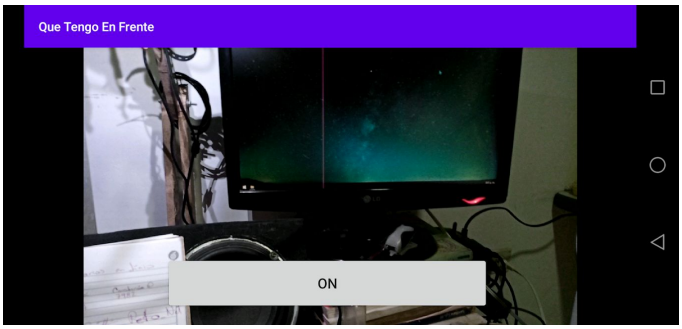

Imagen 1. Prototipo 1 con Boton Pequeño

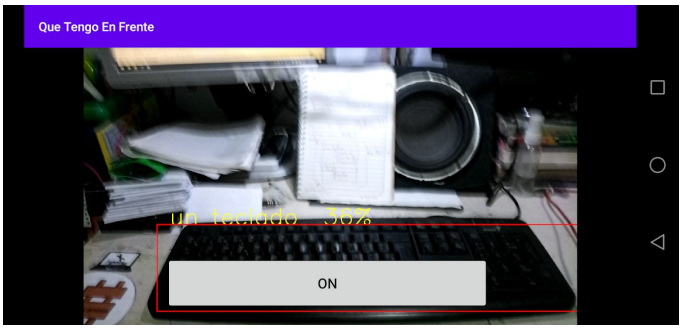

Imagen 2. Ejemplo del reconocimiento de un teclado.

Se estableció un prototipo en el cual una persona con discapacidad visual tendría una mayor facilidad para activar el reconocimiento de objetos.

Imagen 3. Prototipo 2 con Botón más accesible

VI. CONCLUSIONES

En este trabajo se propuso una app identificadora de objetos mediante el uso de imágenes, aplicando una red neuronal que se encargará de responder con el nombre del objeto y el porcentaje de acierto.

Esta app ayudará a las personas con discapacidades visuales a la hora de identificar los objetos que ellos mismos no logren ver con claridad.

Con este trabajo lo que se busca es ayudar a las personas con discapacidad visual a mejorar su calidad de vida de estas personas, a ser un poquito más independientes en su vida.

AGRADECIMIENTOS:

Ivan Goncharov: Nos proporcionó el paso a paso para desarrollar la App y algunos materiales. Puede encontrar estos pasos en el siguiente link:

https://youtu.be/lGtUA5wz0tk?list=PLZBN9cDu0MSl6Ei6OzhVfRtSWaxkt0LA1

VII. REFERENCIAS

[1] E. Sucar and G. Gómez, *Visión Computacional*, 1st ed. Neuherberg, 2020, pp. 1-185.

[2] Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T. M., & An, W. (2018). Detecting nonhardhat-use by a deep learning method from far-field surveillance videos. Automation in Construction, 85, 1-9.

[3] Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. Neural Networks, 1(2), 119–130.

[4] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In European conference on computer vision (pp. 21–37). Springer.

[5] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779–788).

[6] M. Massiris, C. Delrieux and J. Fernández, *DETECCIÓN DE EQUIPOS DE PROTECCIÓN PERSONAL MEDIANTE RED NEURONAL CONVOLUCIONAL YOLO*, 1st ed. 2020, pp. 1-8.

[7] *YOLO9000: Better, Faster, Stronger*, 1st ed. Estados Unidos Seattle, 2017. https://openaccess.thecvf.com/content_cvpr_2017/papers/Redmon_YOLO9000_Better_Faster_CVPR_2017_paper.pdf

[8] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. ArXiv Preprint.

[9] C. Quintero, F. Merchán, A. Cornejo and J. Sánchez, Uso de Redes Neuronales Convolucionales para el Reconocimiento Automático de Imágenes de Macroinvertebrados.

[10] MA Ansari, Krizhevsky, *"Review of Deep Learning Techniques for Object Detection and Classification"*.2018

[11] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).