Universidad Católica de Santa María

Facultad de Ciencias e Ingenierías Físicas y Formales

Escuela Profesional de Ingeniería de Sistemas

RECONOCIMIENTO DE PERSONAS USANDO IMÁGENES DE OREJAS CAPTURADAS EN AMBIENTES NO CONTROLADOS

Tesis presentada por el bachiller:

Ramos Cooper, Solange Griselly

Para optar el Título Profesional de

Ingeniero de Sistemas con especialidad en Sistemas de Información

Asesor**:**

Mg. Montesinos Murillo, Ángel Felipe

Arequipa - Perú

2022

UCSM-ERP

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

INGENIERIA DE SISTEMAS

CON ESPECIALIDAD EN SISTEMAS DE INFORMACION

TITULACIÓN CON TESIS

DICTAMEN APROBACIÓN DE BORRADOR

Arequipa, 26 de Octubre del 2022

Dictamen: 005264-C-EPIS-2022

Visto el borrador del expediente 005264, presentado por:

2010240212 - RAMOS COOPER SOLANGE GRISELLY

Titulado:

RECONOCIMIENTO DE PERSONAS USANDO IMÁGENES DE OREJAS CAPTURADAS EN AMBIENTES NO CONTROLADOS

Nuestro dictamen es:

APROBADO

1568 - ROSAS PAREDES KARINA
DICTAMINADOR

1635 - SULLA TORRES JOSE ALFREDO
DICTAMINADOR

1748 - CALDERON RUIZ GUILLERMO ENRIQUE
DICTAMINADOR

DEDICATORIA

Dedico este trabajo a mi familia.

Agradezco a mis padres, María y René, por su infinito apoyo y confianza, y a mis hermanos, Shirley y Patrick, por su constante aliento.

EPÍGRAFE

"El verdadero progreso se da cuando se pone la tecnología al alcance de todos", *Henry Ford*.

"Me lo contaron y lo olvidé; lo vi y lo entendí; lo hice y lo aprendí", Confucio.

RESUMEN

Las imágenes de orejas como identificadores biométricos son una fuente altamente confiable y estable para el reconocimiento de personas. La oreja humana no sufre cambios drásticos al largo de los años de vida de una persona, no son afectadas por las expresiones faciales o rasgos de envejecimiento y son menos propensas a sufrir daños que alteren su estructura natural. Estas características hacen de las orejas un rasgo biométrico que podría ser usado tanto como los rostros y las huellas dactilares o como un complemento a estos.

En este trabajo se han investigado y analizado diferentes métodos para emplear técnicas de aprendizaje profundo a las tareas involucradas en el reconocimiento de personas usando imágenes de orejas. Las cuales incluyen la detección, la segmentación, y el reconocimiento propiamente dicho. El uso de técnicas de aprendizaje profundo surge de la necesidad de ejecutar el proceso de reconocimiento en ambientes de alta variabilidad y sin un control específico, tales como las situaciones que se dan en interiores y exteriores de centros comerciales, instituciones públicas, privadas, o simplemente en las calles. Este tipo de reconocimiento son usados por sistemas de seguridad o video vigilancia.

Palabras clave:

Orejas, biometría de la oreja, reconocimiento de personas, reconocimiento de orejas, aprendizaje profundo, redes neuronales convolucionales.

ABSTRACT

Ear images as biometric identifiers are a highly reliable and stable source for people recognition. The human ear does not suffer drastic changes throughout a person's lifetime, they are not affected by facial expressions or aging features and are less prone to damage that alters their natural structure. These characteristics make ears a biometric trait that could be used as well as faces and fingerprints or just as a complement to them.

In this work, different methods have been investigated and analyzed to apply deep learning techniques to the tasks involved in the recognition of people using ear images. These tasks include detection, segmentation, and recognition itself. The use of deep learning techniques arises from the need to execute the recognition process in environments of high variability and without specific control, such as situations that occur inside and outside of malls, public or private institutions, or simply on the streets. This type of recognition is used by security or video surveillance systems.

Key words:

Ears, ear biometrics, people recognition, ear recognition, deep learning, convolutional neural networks.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

El reconocimiento de personas es una necesidad constante en la sociedad, esto ha generado que campos como la biometría se conviertan en áreas de investigación activas a lo largo de varios años. Sistemas basados en reconocimiento biométrico son ampliamente usados en aplicaciones de seguridad, video vigilancia y ciencias forenses. Algunos rasgos físicos comúnmente usados en estos sistemas son las huellas dactilares, el rostro y el iris. Aunque inicialmente este tipo de sistemas se utilizaban como herramientas para investigaciones criminales, en la actualidad los sistemas de verificación basados en biometría son bastante comunes y usados para una amplia variedad de propósitos comerciales y de seguridad. Por ejemplo, se utilizan en la banca en línea, el comercio electrónico, la atención médica, el check-in en el aeropuerto, la seguridad fronteriza, el desbloqueo de teléfonos móviles, la aplicación de la ley entre otros.

No obstante, otro rasgo que puede ser usado como identificador biométrico para el reconocimiento de personas son las orejas. Las características anatómicas de las orejas difieren de una persona a otra y han sido reconocidas como un medio de identificación personal. Bertillon (1890), un criminólogo francés, fue el primero en investigar las orejas y su potencial para la identificación de individuos. En 1964, un oficial de policía estadounidense, llamado Alfred Iannarelli, recopiló más de 10,000 imágenes de orejas y realizó algunos estudios en los que determinó 12 características de las orejas para identificar a un individuo (Iannarelli, 1989). Algunas características que hacen que las orejas resulten en una fuente de información confiable y estable son que las orejas no sufren cambios drásticos a medida que la persona envejece y tampoco la estructura es afectada por las expresiones faciales. De acuerdo con Jain, Flynn, & Ross (2010) la estructura de la oreja permanece sin cambios entre la edad de 8 y 70 años, después de los 70 años la oreja crece de forma longitudinal, pero sin alterar su estructura.

Además, las orejas pueden incluso ser distinguidas entre gemelos idénticos, según Nejati, Zhang, Sim, Martinez-Marroquin, & Dong (2012).

El reconocimiento biométrico de la oreja se refiere a la tarea de reconocer personas a partir de imágenes de la oreja utilizando técnicas de visión artificial y aprendizaje automático. Esta tarea ha sido investigada en los últimos años logrando grandes mejoras. Sin embargo, enfoques como la detección de orejas y el reconocimiento de orejas se han abordado de forma independiente sin tener en cuenta que forman parte de un proceso unificado y deben investigarse como un todo. El empleo de técnicas de visión por computadora y aprendizaje automático en este tipo de tareas es debido a que, en la última década, los algoritmos basados en técnicas de aprendizaje profundo y modelos de redes neuronales convolucionales han tenido un buen rendimiento en tareas como clasificación de imágenes y detección de objetos.

A pesar de que las orejas tienen ventajas distintivas sobre otros rasgos físicos, su reconocimiento presenta desafíos importantes que deben ser abordados. Uno de ellos es la oclusión, esta puede ser producida por el cabello, audífonos, auriculares, aretes, piercings, u otros accesorios para las orejas o la cabeza. Así como otros problemas como la variación en la iluminación, la posición de la cabeza, imágenes borrosas, baja resolución, desenfoque, entre otros. Estos problemas generalmente se presentan cuando la captura de imágenes se realiza en condiciones no controladas, lo que se conoce también como reconocimiento en la naturaleza. El reconocimiento en condiciones no controladas hace referencia al hecho de capturar las imágenes en escenarios de la vida real donde existe una alta variación generada por el entorno. Esta variación genera grandes desafíos para las tareas de detección, segmentación y reconocimiento.

CAPÍTULO I

1 PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1 Descripción del problema

1.1.1 Definición del problema

El reconocimiento de personas usando imágenes de orejas capturadas en condiciones controladas ha sido ampliamente explorado en los últimos años, sin embargo, hay pocos experimentos realizados con imágenes tomadas en ambientes no controlados (Kamboj, Rani, & Nigam, 2022) (Booysens & Viriri, Ear Biometrics Using Deep Learning: A Survey, 2022). Además, usualmente se abordan las tareas de detección y segmentación por separado de las tareas de extracción de características y reconocimiento (Cintas, y otros, 2019) (Alshazly, Linse, Barth, & Martinetz, 2020). Es importante abordar y evaluar todas las tareas en conjunto en un proceso end-to-end, ya que el rendimiento de tareas previas tendrá un impacto en el rendimiento de tareas posteriores.

1.1.2 Línea y Sub-línea de Investigación

Inteligencia Artificial e Inteligencia Artificial Aplicada

1.1.3 Tipo y Nivel de Investigación

El tipo de investigación es aplicada, ya que se pretende usar conceptos de biometría, principios de aprendizaje profundo y redes neuronales para la aplicación en el reconocimiento de personas usando imágenes de orejas. Por otro lado, el nivel de investigación es exploratorio, correlacional y experimental, ya que el campo de reconocimiento usando orejas aún se encuentra en etapas iniciales al igual que las técnicas aplicadas para el reconocimiento en ambientes no controlados. Así como también se busca,

mediante experimentos, explorar técnicas que fueron usadas en áreas similares y aplicarlas al área de reconocimiento de orejas.

1.2 Objetivos

1.2.1 General

Reconocer personas usando imágenes de orejas capturadas en ambientes no controlados aplicando técnicas de aprendizaje profundo.

1.2.2 Específicos

i. Caracterizar el estado del arte de la Inteligencia Artificial en tareas específicas de detección, segmentación y reconocimiento de objetos, usando redes neuronales convolucionales.

ii. Implementar y entrenar modelos de redes neuronales convolucionales para la detección, segmentación y reconocimiento de orejas.

iii. Validar la efectividad de los modelos entrenados.

iv. Validar el proceso de reconocimiento de personas usando imágenes de orejas en ambientes no controlados.

1.3 Solución propuesta

Un proceso completo de reconocimiento usualmente implica las tareas de detección, segmentación, alineación, descripción y finalmente el reconocimiento en sí. En tanto, antes del reconocimiento, es necesario abordar las tareas previas de forma efectiva ya que esto tendrá un impacto en las siguientes fases del reconocimiento. En este trabajo se propone utilizar técnicas de aprendizaje profundo o *Deep Learning* (DL) y Redes Neuronales Convolucionales o *Convolutional Neural Networks* (CNN) para abordar los problemas de detección y segmentación, así como también el reconocimiento de imágenes de orejas. La

propuesta implica explorar diferentes modelos, entrenarlos y ajustarlos a las tareas y problemas mencionados.

1.4 Justificación

La investigación de técnicas DL son áreas que vienen siendo exploradas hace varios años, así como su aplicación en distintos problemas y situaciones del día a día. Las CNN han tenido un gran éxito en aplicaciones prácticas y constituyen el estado del arte en muchos problemas de visión computacional, esto se debe al hecho de que han demostrado ser muy eficaces para la clasificación de imágenes a gran escala (Akcay, Kundegorski, Willcocks, & Breckon, 2018) (Liang, He, Li, & Wang, 2019) (Medus, Saban, Francés-Víllora, Bataller-Mompeán, & Rosado-Muñoz, 2021). Además, una de las competencias más conocidas de visión computacional, el *ImageNet Large-Scale Visual Recognition Challenge* (ILSVRC) (Stanford Vision Lab, 2015), evidencia que los algoritmos basados en técnicas de aprendizaje profundo alcanzan los mejores resultados en tareas como detección, localización y clasificación de objetos.

1.5 Alcance y limitaciones

En cuanto al alcance de esta investigación, este abarca la evaluación de los principales modelos propuestos en la literatura que hagan uso de técnicas DL y que hayan sido aplicados en contextos similares como el reconocimiento facial o reconocimiento de otras características biométricas. Mientras que las limitaciones están dadas por la cantidad y disponibilidad de bases de datos propuestas en la literatura para resolver tareas de visión computacional, ya que la cantidad y extensión de los datos son un elemento importante cuando se aplica técnicas de aprendizaje profundo.

CAPÍTULO II

2 FUNDAMENTOS TEÓRICOS

Este capítulo presenta conceptos básicos para una mejor comprensión del trabajo realizado, así como también describe la revisión del estado del arte en cuanto a detección de imágenes de orejas y reconocimiento de personas usando imágenes de orejas. Finalmente, la última sección describe las técnicas generales de entrenamiento de redes neuronales convolucionales y las herramientas usadas para el entrenamiento de los modelos.

2.1 Bases teóricas del Proyecto

2.1.1 Biometría de la Oreja

Recientes investigaciones han demostrado que el reconocimiento basado en la oreja o en la estructura visible del oído externo, no es menos efectiva que otros tipos de reconocimiento, tales como rostro o huellas dactilares (Jain, Flynn, & Ross, 2010). De hecho, el desarrollo de diferentes técnicas ha permitido que el reconocimiento de oreja alcance un rendimiento similar al de reconocimiento basado en imágenes de rostros por ejemplo. Ha sido también experimentalmente probado que las orejas son únicas en cada individuo (Bertillon, 1890) (Iannarelli, 1989). Además, sus características son más estables a lo largo de la vida de una persona, comparado con el envejecimiento de rostros o el daño causado en las huellas digitales debido a tareas repetitivas. Esto contribuye a que las orejas sea un elemento relativamente estable para la identificación y verificación de personas. Además, la adquisición y manejo de imágenes de orejas tiene un bajo costo, ya que pueden ser capturadas con una cámara simple, no se necesita ningún tipo de sensor especializado, y no se necesita la activa cooperación de los individuos. Estas características incrementan aún más el interés en este campo.

La Figura 1 muestra la apariencia de la oreja, la cual está definida por el hélix, antihelix, trago, antitrago, la concha, las fosas triangular y escafoidea, los pilares del antihelix y el lóbulo.

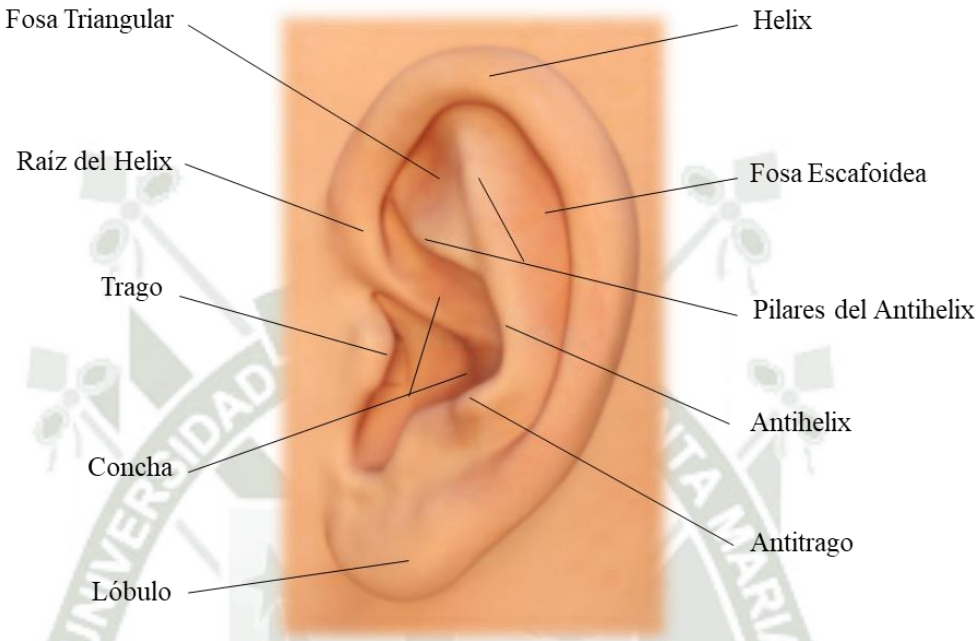

Figura 1 Estructura de la oreja
Fuente: Elaboración propia

A pesar de que el reconocimiento de oreja no es usado ampliamente aún para la identificación y/o verificación de personas, existe concreta evidencia que es adecuado y confiable para ser usado e implementado en sistemas biométricos (Pflug & Busch, 2012) (Jain, Flynn, & Ross, 2010).

2.1.2 Aprendizaje Profundo

El aprendizaje profundo o Deep Learning (DL) es un subconjunto del Aprendizaje Automático o Machine Learning (ML) que se basa en el uso de Redes Neuronales Artificiales. La

Figura 2 muestra la relación entre las técnicas de ML, DL y el área de la Inteligencia Artificial o Artificial Intelligence (AI). Se le llama aprendizaje profundo debido a que las topologías o arquitecturas usan redes neuronales artificiales de múltiples capas, ya sean de entrada, salida u ocultas. Cada capa contiene unidades que transforman los datos de entrada en datos que serán usado por la siguiente capa, por lo tanto, estos algoritmos aprenden cómo hacer una predicción a través de su propio procesamiento de datos (Microsoft Documentation, 2022). A diferencia de las técnicas de ML, a estos algoritmos se les debe proporcionar información extra que debe ser previamente obtenida, por ejemplo, usando técnicas de extracción de características manuales (ver Figura 3).

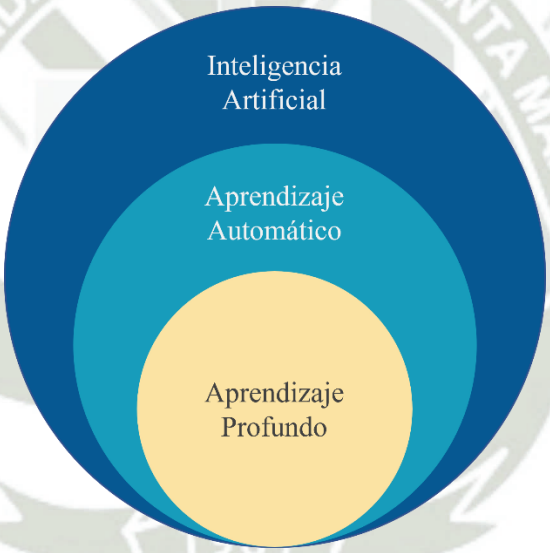

Figura 2 El aprendizaje Profundo y la Inteligencia Artificial

Fuente: Elaboración propia

Gracias a la estructura de las redes neuronales artificiales, las técnicas de aprendizaje profundo son capaces de identificar patrones en datos no estructurados tales como imágenes, sonido, video y texto. Las redes neuronales artificiales están formadas por capas de nodos conectados y para el caso de las técnicas de aprendizaje profundo estas redes contine una gran cantidad de capas. Algunas de las topologías que se puede mencionar dentro de las redes neuronales artificiales son: las redes de retroalimentación o *Multilayer Perceptron*

(MLP), las redes neuronales recurrentes o *Recurrent Neural Networks* (RNN), las redes neuronales convolucionales o *Convolutional Neural Networks* (CNN), las redes adversarias generativas o *Generative Adversarial Network* (GAN) y las redes transformadoras o *Transformers*.

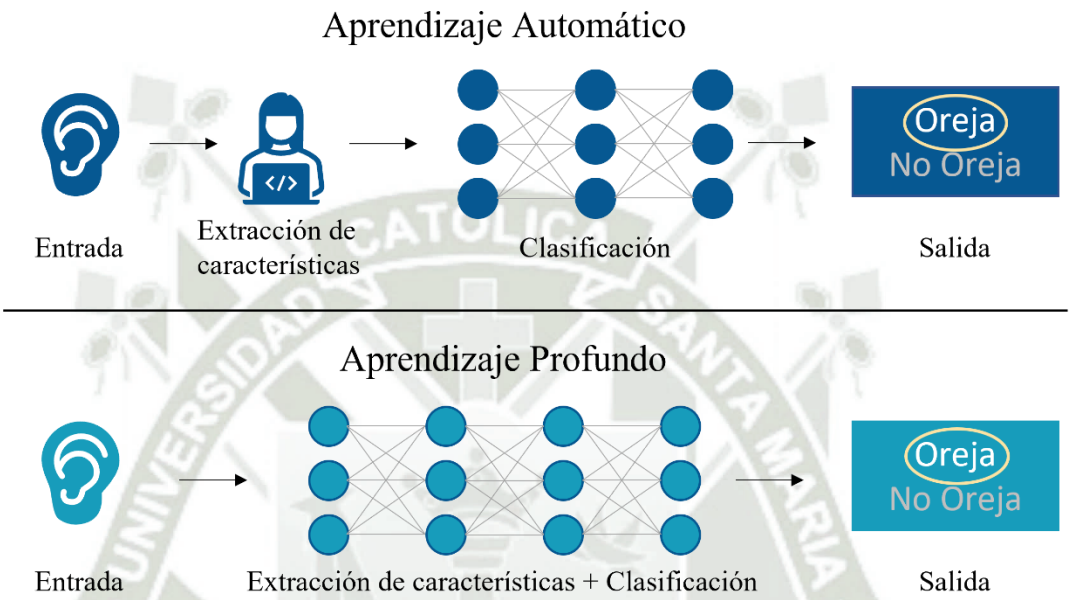

Figura 3 Aprendizaje automático vs. aprendizaje profundo para un caso de clasificación de dos tipos o clases de objetos: oreja y no oreja.

Fuente: Elaboración propia

2.1.3 Redes Neuronales Convolucionales

La red neuronal convolucional es un tipo especializado para el procesamiento de datos que tienen una topología tipo cuadrícula. Un ejemplo claro son los datos de imagen, que pueden pensarse como una cuadrícula 2D de píxeles. El nombre "convolucional" indica que la red emplea una operación matemática llamada convolución que es un tipo especializado de operación lineal (Goodfellow, Bengio, & Courville, 2016). La arquitectura de una CNN típica consiste en dos partes: la parte de extracción de características, que usa operaciones de convolución para extraer características de una imagen; y la parte de clasificación, que es ejecutada haciendo uso de un Perceptrón Multicapa (ver Figura 4).

La parte de clasificación de una CNN se basa en el tradicional modelo de una red neuronal artificial, el cual puede ser definido de la siguiente manera. Dada una red neuronal de *N* capas, sea X_0 y X_N los vectores de entrada y salida respectivamente de la red. Donde el vector X_{n-1} es la entrada de la capa *n* (donde *n* =1, …, *N*). Además, si W_n es una matriz de pesos y b_n un vector de bias, entonces la capa de salida, X_n, puede ser representada con el siguiente vector:

$$X_n = f\ (W_n\ X_{n-1} + \ b_n),$$

donde f es una función de activación. En un problema de reconocimiento, el proceso de entrenamiento consiste en encontrar los valores óptimos para el conjunto de parámetros de pesos y bias $\{W_n, b_n\}$ que minimice el error de clasificación. Es una práctica común utilizar el algoritmo de gradiente descendiente para determinar cómo deben cambiar los parámetros para disminuir el error. La definición del error depende del tipo de dato de salida (Goodfellow, Bengio, & Courville, 2016).

La Figura 4 muestra la arquitectura básica de una CNN usada para clasificación. Esta arquitectura presenta 4 tipos de capas principalmente. Capa convolucional, capa de agrupación o *Pooling*, capa de unidad lineal rectificada o *ReLU*, y capa completamente conectada o *Dense Layer*. Las capas de convolución toman una imagen como entrada y usan una matriz de filtro para realizar la convolución. La capa de agrupación se usa comúnmente para reducir la dimensionalidad y extraer la información más significativa de cada sección de la imagen. Las capas *ReLU* se utilizan para llevar la no linealidad a la red. Finalmente, las capas completamente conectadas se utilizan para conectar todas las neuronas de la capa anterior a la siguiente capa. En una red usada para clasificación, la última capa también se le conoce como capa de clasificación.

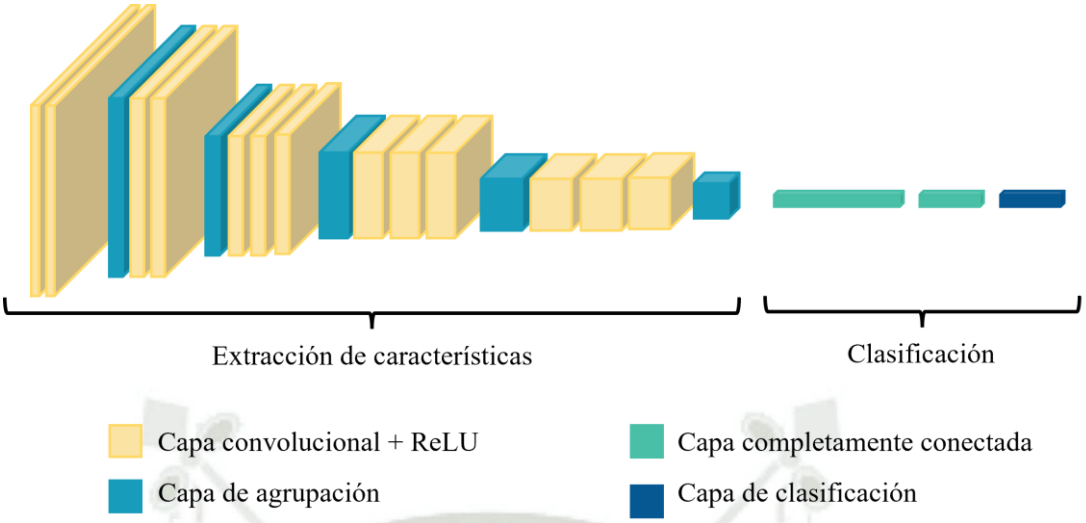

Figura 4 Estructura común para una CNN

Fuente: Elaboración propia

2.1.4 Redes Neuronales Convolucionales Basadas en Regiones para la Detección de Objetos (R-CNN)

Este tipo de arquitectura fue creada para resolver el problema de la detección de objetos, la tarea que consiste en encontrar diferentes objetos en una imagen y clasificarlos. Una CNN basada en Regiones (R-CNN) toma una imagen y genera cuadros delimitadores con la etiqueta correspondiente para cada objeto en la imagen. Los cuadros delimitadores son propuestos usando un proceso llamado búsqueda selectiva, este proceso agrupa píxeles adyacentes por color, textura o intensidad para generar un conjunto de propuestas. Luego, las imágenes de los cuadros delimitadores propuestos pasan por una red previamente entrenada para averiguar qué objeto hay en la imagen. Finalmente, se realiza un proceso de ajuste basado en regresión lineal para mejorar los cuadros delimitadores y obtener coordenadas más ajustadas al objeto. (Girshick, Donahue, Darrell, & Malik, 2014)

2.1.4.1 La Rápida R-CNN o Fast R-CNN

Con el objetivo de mejorar el tiempo de rendimiento de una R-CNN, Girshick (2015) propuso la Rápida R-CNN o *Fast R-CNN*. Una R-CNN genera alrededor de 2,000 propuestas de regiones para cada imagen, las cuales pasan a través de otra CNN para extraer características, dando como resultado 2,000 pases por imagen. Esto agregado al clasificador que predice la clase y el modelo de regresión que ajusta los cuadros delimitadores, hacía que el modelo fuera realmente lento para entrenar. La *Fast R-CNN* resuelve este problema mediante el uso de una técnica conocida como agrupación de regiones de interés o *RoIPool*. Esta técnica comparte el mapa de características de una imagen entre sus subregiones, por lo que cada subregión toma su correspondiente región del mapa de características para luego aplicar una agrupación máxima para finalmente obtener el mapa de características para cada subregión. De esta forma, las 2,000 pasadas que se realizaban por imagen se reducen a una sola. Además, la *Fast R-CNN* une la red convolucional, el clasificador y el regresor de cuadro delimitador que funcionaban por separado en el modelo R-CNN. Por lo tanto, la Fast R-CNN usa una sola red para generar los dos resultados, cuadros delimitadores y clasificación.

2.1.4.2 La Más Rápida R-CNN o Faster R-CNN

En el año 2017, se publicó un trabajo que proponía hacer esta arquitectura aún más rápida que la propuesta anterior. Los autores descubrieron que el algoritmo de búsqueda selectiva que proponía las regiones que sirven como entrada a la R-CNN era la parte que hacía que todo el proceso fuera muy lento. Ellos observaron que el mapa de características convolucionales utilizado por la Rápida R-CNN también se puede usar para generar propuestas de región. De esta manera, en lugar de usar un algoritmo externo para proponer las regiones, la más rápida R-CNN posee una CNN extra que obtiene las regiones de propuesta durante el entrenamiento (Ren, He, Girshick, & Sun, 2017).

2.1.4.3 La Máscara R-CNN o Mask R-CNN

Algunos años más tarde, la detección de objetos basada en cuadros delimitadores evolucionó a una segmentación a nivel de píxel con la propuesta *Mask R-CNN* presentada por un grupo de investigación en AI de Facebook (He, Gkioxari, Dollar, & Girshick, 2017). La *Mask R-CNN* aborda la segmentación agregando una rama a la *Fast R-CNN* que genera una máscara binaria, es decir una matriz de unos y ceros, los 1 son ubicados en píxeles que pertenecen al objeto y los 0 en los demás pixeles.

La gran diferencia de la Mask R-CNN con respecto a sus predecesores es que no solo puede realizar la detección de objetos, sino también la segmentación de instancias. Esto significa que el algoritmo puede predecir los cuadros delimitadores para todos los objetos de una imagen que pertenecen a una clase, y también puede especificar todos los píxeles que pertenecen a cada objeto detectado en la imagen. Figura 5 muestra la diferencia entre cuatro tareas similares de visión artificial: clasificación, segmentación semántica, detección de objetos y segmentación de instancias, todas ellas realizadas sobre la misma imagen. Un algoritmo de clasificación puede predecir que hay un objeto que pertenece a la clase “globo”. Una segmentación semántica puede generar todos los píxeles que pertenecen a todos los objetos de la clase “globo”. Un algoritmo de detección de objetos, como la Más Rápida R-CNN, puede predecir la ubicación de los 7 objetos globo. Aunque los objetos se superponen, la predicción puede especificar cuadros delimitadores para cada objeto globo. Finalmente, un algoritmo de segmentación de instancias, como la Mask R-CNN, predecirá la ubicación de los 7 objetos y los píxeles que pertenecen a cada objeto, marcando la diferencia para cada instancia. En otras palabras, muestra la ubicación del primer globo y sus correspondientes píxeles, la ubicación del segundo globo y sus correspondientes píxeles, y así sucesivamente para las 7 instancias de objetos de clase globo.

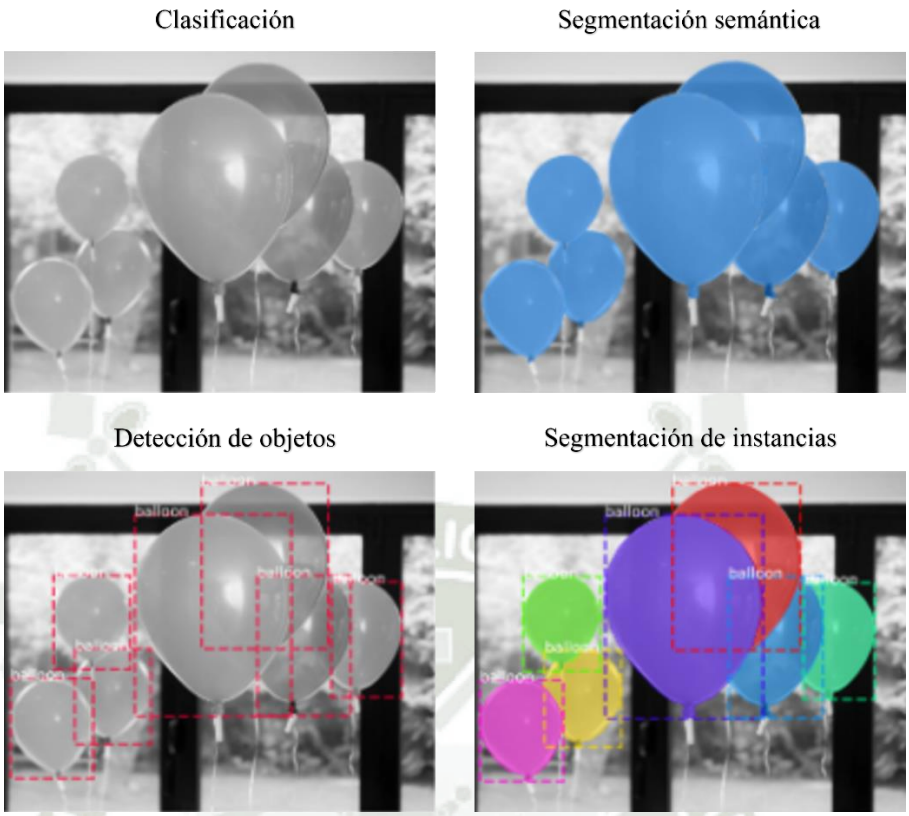

Figura 5 Detección y segmentación de objetos

Fuente: Imagen extraída de Abdulla (2019)

Dado que la detección y segmentación de objetos es la primera etapa en el proceso reconocimiento, su rendimiento tiene una gran influencia en todo el sistema de reconocimiento. Por lo tanto, agregar una CNN basada en segmentación de instancias que detecte oídos a nivel de píxel podría mejorar el rendimiento de la próxima etapa, así como el proceso de reconocimiento completo.

2.1.5 Redes Neuronales Convolucionales para Reconocimiento de Objetos

La tarea de extracción de características mediante imágenes implica extraer un mayor nivel de información de los valores de píxeles de la imagen que pueden capturar la distinción entre categorías. Se han explorado ampliamente diferentes técnicas que utilizan modelos DL pre-entrenados debido a que las bibliotecas y frameworks como Tensorflow (Google Brain Team, 2015) y PyTorch (Facebook's AI Research lab (FAIR), 2016) contienen estos

modelos básicos pre-entrenados en algunos conjuntos de datos a gran escala conocidos, como ImageNet (Stanford Vision Lab, Stanford University, 2020) o Coco (Lin, y otros, 2015). El modelo utilizado para la extracción de características en este trabajo se describe a continuación.

2.1.5.1 VGG

Este modelo fue propuesto por Simonyan & Zisserman (2015) de la Universidad de Oxford y fue presentado en la competencia del Image Large-Scale Visual Recognition Challenge en el año 2014. Este modelo alcanzó un 92,7% de precisión en la prueba top-5 sobre 14 millones de imágenes pertenecientes a 1000 clases. Los autores investigaron la profundidad de los modelos y su impacto en el reconocimiento de imágenes a gran escala. Descubrieron que al aumentar la profundidad y usar filtros muy pequeños logran una mejora significativa en la tarea de reconocimiento. El tamaño de la imagen de entrada en este modelo es de 224 x 224 píxeles. Los filtros son de tamaño de 3 x 3 para capturar la noción de izquierda/derecha, arriba/abajo, centro. El paso de convolución se fija en 1 píxel para conservar la resolución espacial después de la convolución. La agrupación espacial se lleva a cabo mediante cinco capas de agrupación máxima, que siguen a algunas de las capas convolucionales. La agrupación máxima se realiza en una ventana de píxeles de 2 x 2, con un paso de 2. La pila de capas convolucionales y de agrupación máxima son seguidas por tres capas completamente conectadas. Las dos primeras con tamaño de 4096 unidades, y la tercera configurada con 1,000 unidades para realizar la clasificación de los 1000 objetos en el conjunto de datos. La capa final utiliza una función de activación llamada *softmax*, la cual es una función matemática que convierte un vector de números en un vector de probabilidades, donde las probabilidades de cada valor son proporcionales a la escala relativa de cada valor en el vector. Estas probabilidades luego son interpretadas como posibles etiquetas para el objeto en la imagen.

2.2 Estado del Arte

2.2.1 Bases de datos de orejas

Esta sección describirá algunos conjuntos de datos de orejas que han sido propuestos y utilizados para entrenar y probar diferentes enfoques en la literatura. Los primeros dos conjuntos abordan el problema de segmentación y presentan anotaciones a nivel de píxel indicando si un píxel pertenece al objeto oreja o no. El resto de los conjuntos se encuentran anotados por sujeto; por lo tanto, se utilizan para problemas de reconocimiento y verificación. Según el tipo de imágenes que recolectan, estos conjuntos pueden clasificarse como conjuntos restringidos, semi-restringidos y no restringidos. Finalmente, el último conjunto de datos fue creado para resolver el problema de alineación y normalización de imágenes de orejas, ya que cuentan con anotaciones de puntos de referencia en la oreja.

2.2.1.1 UBEAR – University of Beira Interior Ear Dataset

Este conjunto reúne 4,430 imágenes en escala de grises de 126 sujetos. La mayoría de los individuos tenían menos de 25 años. Este conjunto de datos no presenta imágenes recortadas alrededor de la oreja, sino que las imágenes de rostro completo (Figura 6). Todas las imágenes tienen el mismo tamaño de 1,280 x 960 píxeles. Además, presentan una máscara binaria con la oreja segmentada para cada imagen, que fue anotada de forma manual. Todas las imágenes se tomaron de grabaciones de video a 3 metros de distancia de la cámara. Luego, se pidió a los individuos que movieran la cabeza hacia arriba y abajo y hacia afuera y adentro. También se les pidió a los sujetos dar un paso hacia adelante y hacia atrás desde la posición inicial. Para cada persona, se capturaron ambos oídos en dos sesiones diferentes, dando un total de cuatro videos por sujeto. Cada secuencia de video fue analizada manualmente, seleccionando alrededor de 17 cuadros para cada persona. Este conjunto de datos también presenta anotaciones de sujetos y anotaciones de oreja izquierda y derecha. Las imágenes fueron almacenadas en formatos TIFF y BMP.

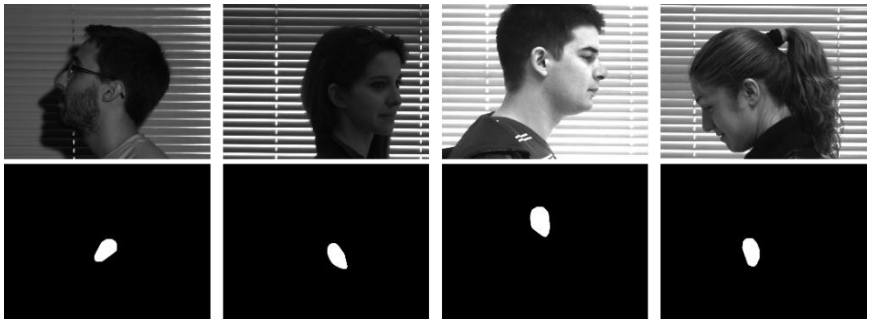

Figura 6 Imágenes de muestra del conjunto de datos UBEAR

Fuente: Raposo, Hoyle, Peixinho, & Proença (2011)

2.2.1.2 AWE - Annotated Web Ears para Segmentación

Este conjunto recopila imágenes de la web tomadas en entornos sin restricciones y de diferentes resoluciones. Contiene 1,000 imágenes con una gran variedad en apariencia, etnia, oclusión, posición de la cabeza, lado de la cabeza e iluminación (Figura 7). El conjunto está dividido en dos grupos: el de entrenamiento con 750 imágenes y el de prueba con 250 imágenes. Cada imagen tiene un tamaño de 480 x 360 píxeles, a color y guardadas en formato PNG. Cada imagen tiene un cuadro delimitador y una imagen de máscara binaria para cada oreja presente en la imagen. No contienen anotaciones de sujeto.

Figura 7 Imágenes de muestra del conjunto de datos AWE para segmentación

Fuente: Emersic Z. , Gabriel, Struc, & Peer (2018)

2.2.1.3 IITD Ear Dataset - Indian Institute of Technology of Delhi

Este es un conjunto restringido que reúne 493 imágenes pertenecientes a 125 sujetos. Cada persona tiene de 3 a 5 muestras de imágenes de la oreja derecha. Las imágenes están en escala de grises, tienen el mismo tamaño 272 x 204 píxeles y están guardadas en formato BMP. Las imágenes en este conjunto no presentan mucha variación en la pose ya que todas las imágenes fueron tomadas en posición de perfil, tampoco en el uso de accesorios y los casos de oclusión son mínimos. Las imágenes se adquirieron usando una cámara digital en un ambiente interior durante nueve meses sin iluminación adicional. Cada voluntario se sentó en una silla y se fijó una cámara digital para adquirir la región de interés, asegurando la presencia de la oreja en las imágenes. Los sujetos tenían 14 y 58 años. La Figura 8 presenta muestras de 3 personas encontradas en este conjunto.

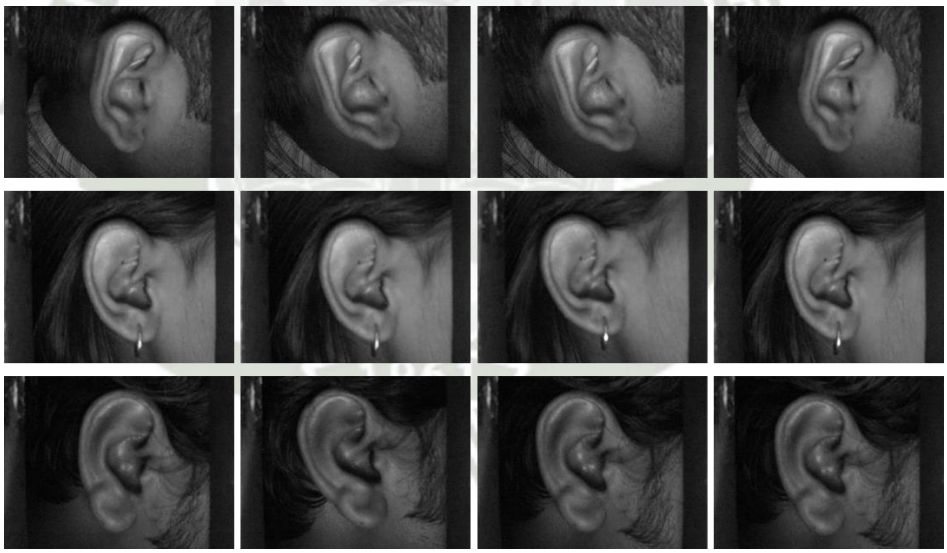

Figura 8 Imágenes de muestra del conjunto de datos IITD
Fuente: Kumar & Wu (2012)

2.2.1.4 AMI - The Mathematical Analysis of Images Dataset

Este conjunto restringido recopila imágenes de estudiantes, profesores y personal del Departamento de Informática de la Universidad de Las Palmas de Gran Canaria en España. Contiene 700 imágenes pertenecientes a 100 personas con edades entre los 19 y 65 años.

Cada sujeto presenta siete muestras, una de la oreja derecha y seis de la oreja izquierda, algunos ejemplos pueden observarse en la Figura 9. Todas las imágenes son a color, tienen el mismo tamaño de 492 x 702 píxeles y están guardas en formato PNG. Las imágenes se tomaron en un ambiente interior bajo las mismas condiciones de iluminación para todos los sujetos. El sujeto se encuentra a unos dos metros de la cámara y mirando unas marcas previamente fijadas, por lo tanto las imágenes presentan poca variación en la iluminación y posición de la cabeza, pero en algunos casos está presente la oclusión y los accesorios.

Figura 9 Imágenes de muestra del conjunto de datos AMI
Fuente: Gonzalez, Alvarez, & Mazorra (2019)

2.2.1.5 WPUT Dataset - The West Pomeranian University of Technology

Este conjunto reúne 2,070 imágenes de 501 individuos (254 mujeres y 247 hombres), la mayoría de ellos en el rango de 21 a 35 años. Cada sujeto presenta de 4 a 8 imágenes (al menos dos imágenes por oreja). Para 451 sujetos, las muestras se tomaron en una sola sesión, mientras que el resto tuvo sesiones repetidas. Este conjunto de datos se creó para reflejar las condiciones de la vida real y recopilar la mayor variedad posible. Estas variaciones están

asociadas a la oclusión provocada por cabello, aretes, accesorios de orejas, anteojos, suciedad, polvo, accesorios para la cabeza, entre otros. El 20% de las fotos están libres de oclusiones, el 15,6% fueron tomadas al aire libre, el 2% fueron tomadas en la oscuridad. A pesar de que las imágenes muestran una amplia gama de variabilidad de apariencia, la gran mayoría aún se adquirieron en entornos similares a los de un laboratorio (Ver Figura 10). Por lo tanto, este conjunto se podría clasificar como un conjunto de datos semi-restringido. Todas las imágenes son a color, con el mismo tamaño 380 x 500 píxeles y se guardan en formato JPG.

Figura 10 Imágenes de muestra del conjunto de datos WPUT

Fuente: Frejlichowski & Tyszkiewicz (2010)

2.2.1.6 AWE – Annotated Web Ears para Reconocimiento

Este conjunto presentado por Emeršič, Štruc, & Peer (2017) contiene 1,000 imágenes de 100 sujetos, con 10 imágenes por sujeto. Diferente de otros conjuntos, las imágenes de este grupo fueron recopiladas usando web crawlers. Las imágenes incluyen información de género, etnia, accesorios, oclusiones, inclinación y balance de la cabeza, lado de la cabeza, punto del trago de la oreja y un número de indetificación para cada sujeto. Las imágenes

son a color, en formato PNG y los datos están anotados en un archivo JSON. Los mismos autores de este conjunto hicieron una extensión a este conjunto y lo llamaron Annotated Web Ears Extended (AWEx) y fue presentado por Emersic, Meden, Peer, & Struc (2018). Esta extensión reúne 246 clases más, con un total de 4,104 imágenes. Después de algunos años fue ampliado nuevamente y adaptado para el Unconstrained Ear Recognition Challenge (UERC) (Emersic, y otros, 2019) , recopilando 11,804 imágenes de 3,706 sujetos. En conjunto se divide en dos partes, el conjunto de entrenamiento con 2,304 imágenes de 166 sujetos (al menos diez imágenes por sujeto) y el conjunto de prueba con 9,500 imágenes pertenecientes a 3,540 sujetos (con un número variable de imágenes por sujeto). El tamaño de la imagen más pequeña en este conjunto es de 15 x 29 píxeles, y la imagen más grande tiene un tamaño de 557 x 1,008 píxeles. Por otro lado, en el grupo de prueba la imagen más pequeña tiene un tamaño de 11 x 21 píxeles, y la imagen más grande alcanza un tamaño de 499 x 845 píxeles. Esta colección muestra variabilidad no solo entre sujetos sino también entre imágenes de cada sujeto (Ver Figura 11). En algunos casos, la diferencia en el momento en que se tomaron las imágenes de un individuo es de varios años de diferencia, algo que no se puede encontrar en conjuntos de datos construidos en laboratorios restringidos. Así como también las imágenes muestran alta variabilidad en la iluminación, posición de la cabeza, oclusión, origen étnico y edad.

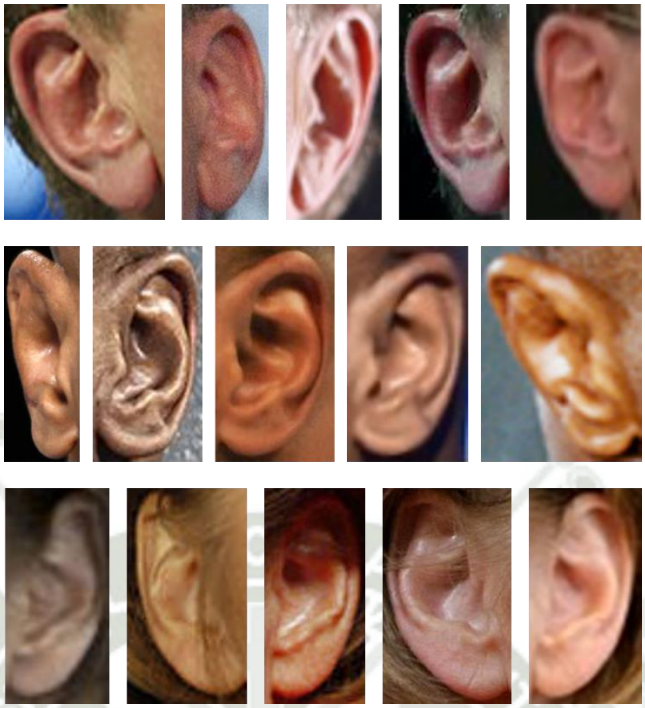

Figura 11 Imágenes de muestra del conjunto de datos AWE para reconocimiento

Fuente: Emersic (2019)

2.2.1.7 EarVN1.0

Este conjunto reúne 28,412 imágenes de 164 sujetos. Esta colección sería etiquetada como un conjunto de datos sin restricciones ya que las imágenes presentan una alta variabilidad entre las muestras de una clase con respecto a la iluminación, la posición de la cabeza, la escala, el contraste, la oclusión y el uso de accesorios. No obstante, todas las imágenes pertenecen a personas de la misma región y etnia, lo que podría generar un sesgo al entrenar un modelo. Cada clase tiene al menos 100 muestras de ambas orejas. Todos los sujetos fueron voluntarios, 98 hombres y 66 mujeres. Las imágenes faciales originales se adquirieron en un entorno sin restricciones de un sitema de cámaras y con condiciones de luz variables. Entre todas las imágenes, la imagen más pequeña tiene un tamaño de 13 x 17 píxeles, y la imagen más grande alcanza un tamaño de 436 x 581píxeles. Las muestras se guardan en formato JPEG. Algunos ejemplos encontrados en este conjunto de datos pueden observarse en la Figura 12.

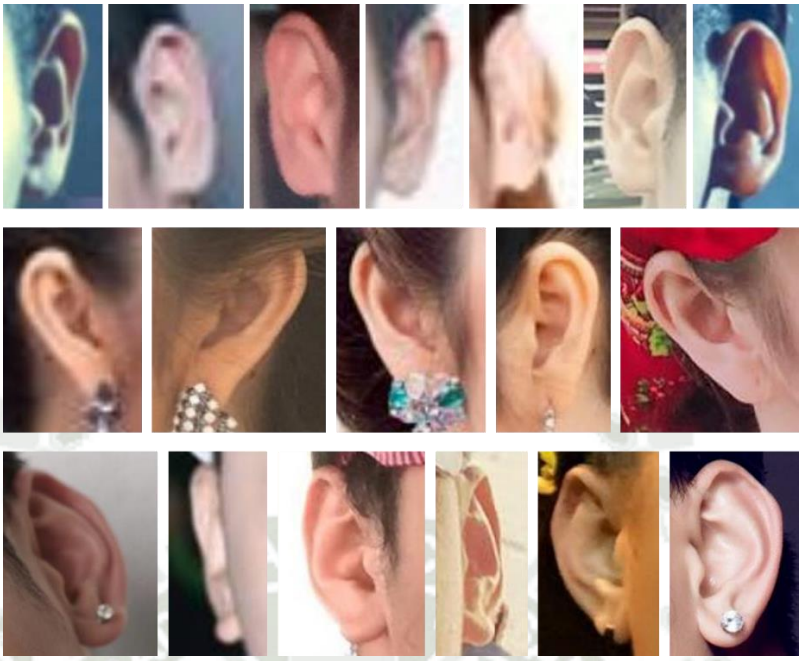

Figura 12 Imágenes de muestra del conjunto de datos EarVN1.0

Fuente: Vinh (2019)

2.2.1.8 VGGFace-Ear

Este conjunto de datos es el más grande que se ha podido encontrar en la literatura al momento de escribir este informe. Este conjunto fue construido como una extensión de una base de datos de rostros llamada VGGFace y presentado por Cao, Shen, Xie, Parkhi, & Zisserman (2018). Este conjunto de datos contiene 234,651 imágenes pertenecientes de 660 sujetos, las clases presentan entre 150 y 645 imágenes de muestra. El conjunto está dividido en dos subconjuntos: el conjunto de entrenamiento que contiene 600 clases o identidades y el conjunto de prueba que contiene 60 clases de 150 muestras por clase. Las imágenes son de tamaño variable, se pueden encontrar imágenes desde dimensiones de 15x15 pixeles hasta más de 650x650 píxeles. Es importante mencionar que las imágenes originales fueron recolectadas de la web y presentan alta variabilidad en cuanto a la rotación de la cabeza, iluminación y oclusión por accesorios o cabello, distancia desde la cámara, entre otras características encontradas en imágenes tomadas en ambientes no controlados. La Figura 13 presenta algunas imágenes encontradas en este conjunto de datos, cada fila presenta seis

muestras de una misma persona, puede notarse la variabilidad entre los ejemplos de cada persona y la variabilidad entre los rasgos de diferentes personas.

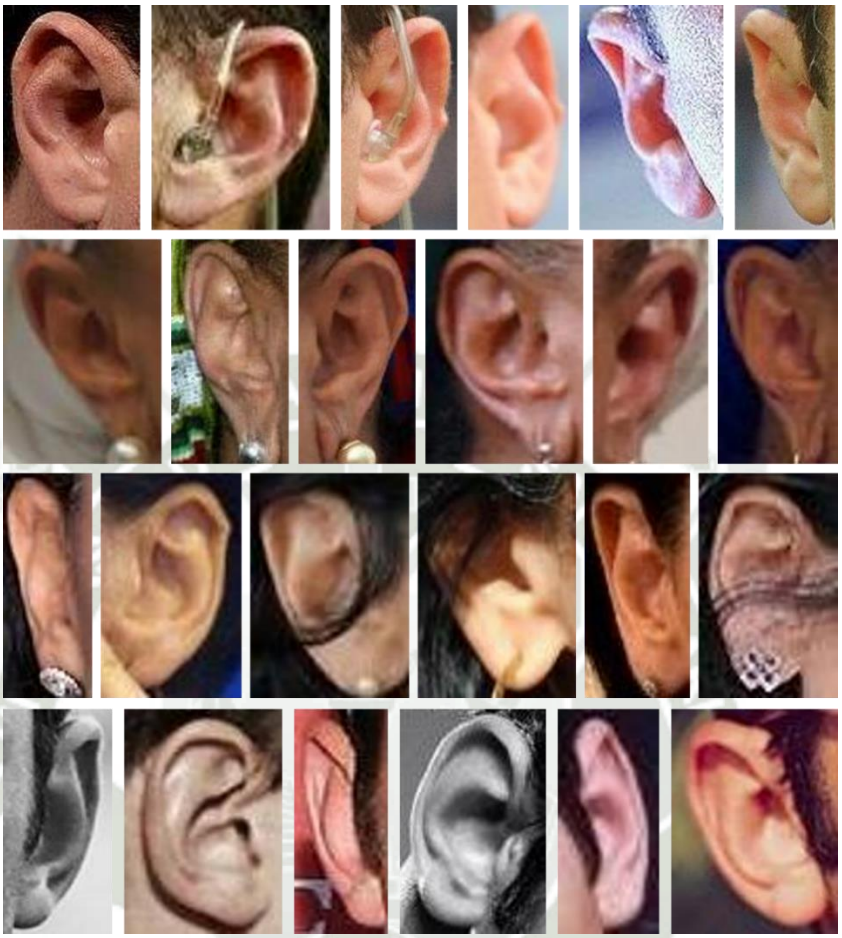

Figura 13 Imágenes de muestra del conjunto de datos VGGFace-Ear

Fuente: Ramos-Cooper, Gomez-Nieto y Camara-Chavez (2022)

2.2.1.9 In The Wild Ear Dataset

Este conjunto reúne 2,058 imágenes pertenecientes a 231 personas, donde todos ellos son personajes públicos. Este conjunto de datos contiene dos colecciones, la primera colección, llamada A, reúne 605 imágenes. Este conjunto se dividió aleatoriamente en dos conjuntos, un conjunto de entrenamiento con 500 clases y un conjunto de prueba con 105 clases. La segunda colección, llamada B, tiene anotaciones de puntos de referencia y anotaciones de clase con el nombre de la persona en la imagen, que contiene de 1 a 14 imágenes por clase.

Cabe mencionar que ambas colecciones incluyen las mismas 2,058 imágenes, pero están agrupadas de manera diferente. Todas las imágenes son de tamaño variable, guardadas en formato PNG y con un archivo de texto plano con las anotaciones de los puntos de referencia. La imagen más pequeña tiene un tamaño de 91 x 288 píxeles, mientras que la más grande tiene un tamaño de 1200 x 1600 pixels. Cada imagen se anotó manualmente con 55 puntos de referencia.

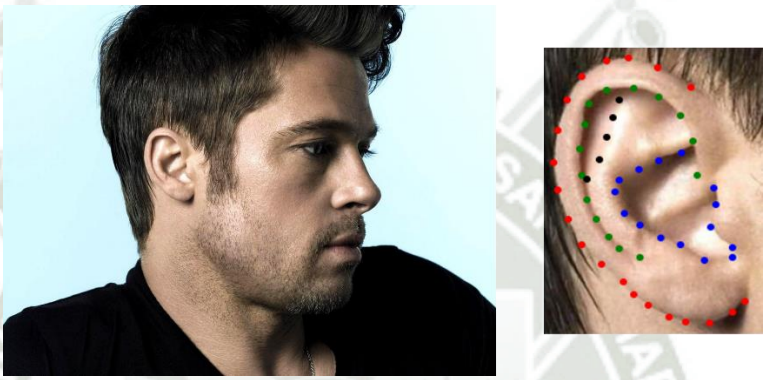

Figura 14 Anotación de puntos de referencia en las imágenes del conjunto de datos In The Wild Ear Dataset
Fuente: Zhou y Zaferiou (2017)

2.2.2 Detección de Orejas

En esta sección, se hará una revisión de los últimos enfoques y técnicas propuestas en la tarea de detección de orejas, con especial atención a los trabajos que se han probado en entornos no controlados y condiciones variables con imágenes 2D.

Algunas de las primeras propuestas en la literatura se basan en técnicas como la detección de bordes, la comparación de plantillas, la transformación de distancias, los contornos activos e incluso algunos que usan información de textura y color, entre otras técnicas. Muchas de estas técnicas fueron ampliamente exploradas en trabajos presentados en Ansari & Gupta (2007), Yuan & Mu (2007), Attarchi, Faez, & Rafiei (2008), Prakash, Jayaraman, & Gupta (2008), Prakash & Gupta (2012), Chidananda, Srinivas, Manikantan, & Ramachandran (2015), Deepak, Vasant, & Manikantan (2016) y Halawani & Li (2016). No

obstante, este tipo de métodos por lo general no resultan lo suficientemente robustos en condiciones altamente variables causando una caída en las tasas de detección (Zhang & Mu, 2017).

Las CNN son técnicas de DL que han sido ampliamente utilizadas en tareas de visión por computadora tales como clasificación de imágenes y detección de objetos. Así que también comenzaron a ser protagonistas en la tarea de detección de orejas.

Un primer método para construir un detector de orejas usando una CNN es el trabajo presentado por Wang, Du, & Huang (2017). Los autores ajustaron la red AlexNet para entrenar un clasificador de orejas, reuniendo una gran cantidad de muestras positivas y negativas para este entrenamiento. Luego, convirtieron las capas completamente conectadas en capas completamente convolucionales e hicieron uso de un enfoque de ventana o cuadro deslizante para construir el detector de oído. Reportaron tasas de detección de alta precisión; sin embargo, esta tasa varía según la cantidad de muestras positivas que utilizan, cuantas más muestras, mayor precisión pueden alcanzar.

Por otro lado, un método que, en lugar de usar un enfoque basado en ventanas deslizantes, usa un enfoque basado en regiones es el método presentado por Zhang & Mu (2017). Los autores presentaron un método para recortar la región de la oreja combinando el contexto de ubicación de la oreja y las características morfológicas, concluyendo que la información del contexto mejora el rendimiento de la detección. El método hace uso de una CNN basada en regiones de múltiples escalas que reconoce tres regiones en una imagen: cabeza, región pan-oreja y oreja. Los autores informan tasas de precisión que fueron evaluados en los conjuntos de datos WebEars, UND-J2 y UBEAR, alcanzando una precisión de 98%, 100% y 98,7% para cada conjunto de datos, respectivamente.

Ganapathi, Prakash, Dave, & Bakshi (2020) se presentó un enfoque diferente usando un conjunto de tres CNNs ensambladas para detectar orejas en imágenes de rostros de perfil, sin estar limitado a un solo rostro. El modelo funciona en dos partes, la primera parte de la técnica entrena tres modelos de CNN en un conjunto de datos dado, luego una parte posterior obtiene un promedio ponderado de la salida de los modelos entrenados para detectar regiones de la oreja. Los autores concluyen que un modelo ensamblado supera a los modelos basados en una sola CNN. La técnica se validó en dos conjuntos de datos: en el IIT colección A y el AWE alcanzando 98,2% y 99,5% de precisión, respectivamente. De la misma manera, Raveane, Galdamez, & González (2019) presentó un método que involucra múltiples CNN para identificar la presencia y ubicación de las orejas en una imagen de entrada dada. Aunque no alcanzan una precisión de alta tasa, su método se probó principalmente en datos de video que reflejan escenarios del mundo real y aplicaciones en tiempo real. Para estos dos últimos enfoques, el principal problema a abordar en futuros trabajos es la disminución de las tasas de falsos positivos. Además, es importante tener en cuenta que todos los enfoques presentados anteriormente realizan una detección de oído de cuadro delimitador, es decir, el resultado de estos métodos son regiones rectangulares que contienen las orejas y, en algunos casos, estas regiones no encajan firmemente en la oreja y agregan alguna información no deseada o innecesaria a la región de interés que puede afectar procesos posteriores como el reconocimiento.

A continuación, se describen diferentes enfoques que ejecutan la detección de orejas de una manera diferente a la generación de regiones de cuadros delimitadores; en su lugar, realizan una localización de oídos a nivel de píxel. Todos estos trabajos se evaluaron utilizando anotaciones por píxeles a través de la métrica de rendimiento de Intersección sobre Unión (IoU), que representa una relación entre la cantidad de píxeles que están presentes en la

intersección de las regiones de las orejas detectadas y las anotadas, sobre la cantidad de píxeles que están en la unión de ambas regiones.

Un trabajo inicial presentado por Emersic Z. , Gabriel, Struc, & Peer (2018) proponía una red de codificador-decodificador convolucional con una tarea de segmentación semántica de dos clases, donde el objetivo es asignar cada píxel de imagen a la clase oído o no oído. La técnica se probó en imágenes 2D de entornos no controlados, pero no funcionan bien en tantas condiciones variables. El modelo alcanza los 55,7% con respecto a la métrica de IoU. Posteriormente, un enfoque de detección presentado en (Emersic, Krizaj, Struc, & Peer, 2019) como parte de un proceso completo de reconocimiento de orejas basado en CNNs, presentó una CNN basada en un modelo existente llamado RefineNet. Los autores validaron la técnica en 250 imágenes pertenecientes al conjunto de datos AWE y obtuvieron un valor de IoU de 84.8\%.

Un trabajo posterior que hace uso del modelo Mask R-CNN (He, Gkioxari, Dollár, & Girshick, 2017) fue presentado por Bizjak, Peer, & Emersic (2019). La arquitectura Mask R-CNN se propuso inicialmente como un detector de objetos de cuadro delimitador con segmentación de instancias en un conjunto de datos a gran escala. En el trabajo de Bizjak, Peer, & Emersic (2019), se utilizó el conjunto de datos de AWE para entrenar y probar esta arquitectura en las tareas de detección de oídos y segmentación de instancias, alcanzando valor de 79,24% IoU. El principal problema de este modelo es que en algunos casos además de objetos de orejas, otros objetos como ojos, lentes, narices y manos también se son detectados como orejas; siendo un problema para la próxima etapa, de reconocimiento.

Finalmente, se evalúa un enfoque diferente que incluye métodos de Morfometría Geométrica o *Geometric Morphometrics* (GM) en (Cintas, y otros, 2019). Este grupo de métodos GM se usa ampliamente en estudios de organismos biológicos. Los autores propusieron cuantificar la forma de cada espécimen de acuerdo con la ubicación en el

espacio de un conjunto de puntos de referencia 2D o 3D que son homólogos entre individuos (Mitteroecker & Gunz, 2009). Los autores entrenaron una CNN con una colección de imágenes anotadas con puntos de referencia basadas en GM para detectar 45 puntos de referencia en las imágenes de oreja, luego usaron un proceso de casco convexo para obtener el conjunto de píxeles correspondientes solo a la estructura de la oreja. Su método se probó con un conjunto de datos utilizado originalmente para el reconocimiento facial y tomado en entornos controlados y entornos restringidos. Luego el método fue comparado con el detector de oído basado en Haar disponible en la biblioteca OpenCV (Bradski, 2000). Por lo tanto, este método no se probó en ningún conjunto de datos de reconocimiento de oído sin restricciones.

2.2.3 Reconocimiento de Oreja

En cuanto al reconocimiento de oreja, existen técnicas tanto tradicionales o artesanales como basadas en DL (Abdulla, 2019). Sin embargo, es importante mencionar que hace solo unos años, los métodos basados en descriptores tradicionales eran lo más avanzado en el campo. Por un lado, las técnicas basadas en descriptores extraen información de áreas locales de la imagen y la utilizan directamente para la identificación; por otro lado, los métodos basados en DL extraen información de la imagen de entrada completa y aprenden la representación de la imagen directamente de los datos de entrenamiento (Booysens & Viriri, 2022). En los siguientes párrafos, se presenta una descripción general de los últimos enfoques sobre el reconocimiento de oreja. La mayoría de los enfoques revisados emplearon modelos de reconocimiento basados en DL.

Cintas y otros (2017) presentaron un primer trabajo que se enfoca en la descripción de características como un paso importante para la etapa de reconocimiento. Los autores han propuesto un método basado en GM y DL para la detección y descripción automática de la oreja. Usando un conjunto de datos a gran escala anotados con puntos de referencia,

entrenaron un modelo para detectar automáticamente 45 puntos de referencia en imágenes 2D en escala de grises y luego usaron estos puntos como un vector de características para la descripción de la oreja. Aunque muestran buenos resultados detectando puntos de referencia, las imágenes de entrada deben estar alineadas antes de entrar en el modelo ya que el modelo no es invariante a la rotación. El conjunto de datos usado en este trabajo es propiedad del Consorcio para el Análisis de la Diversidad y Evolución de los Latinoamericanos (CANDELA, 2016) que contiene 7500 imágenes de la vista lateral de la cabeza sin condiciones específicas de iluminación y sin remoción de fondo. No obstante, este conjunto de datos no se encuentra disponible para fines de investigación.

En lugar de la detección de puntos de referencia, otros enfoques utilizan modelos CNN para aprender a representar las orejas. Sin embargo, uno de los principales problemas cuando se aplican técnicas DL es la poca disponibilidad de datos de entrenamiento.

Un primer trabajo presentado por Emersic, Stepec, Struc, & Peer (2017) empleó un proceso de aumento de datos agresivo para mostrar el impacto en modelos de aprendizaje de tipo completo y selectivo. El aprendizaje completo se refiere a aprender desde cero, mientras que el aprendizaje selectivo se refiere a inicializar un modelo con parámetros ya aprendidos de un conjunto de datos como ImageNet y luego ajustar solo ciertas capas. Los autores usaron tres modelos en sus experimentos VGG16, AlexNet y SqueezeNet. Usaron 1,383 imágenes de oreja tomadas en ambientes no controlados de los conjuntos de datos AWE y CVLED para entrenar, mientras que se hicieron pruebas en un conjunto de 921 imágenes. Entrenar un modelo sin usar aumento de datos resultó en un valor de 41.26% en Rank-1, mientras que, aplicando un aumento de datos en un factor de 100, se alcanzó un valor de 61,89%, empleando un aprendizaje selectivo de SqueezeNet.

Un segundo trabajo presentado por Eyiokur, Yaman, & Ekenel (2018) usa transferencia de aprendizaje para resolver el problema de la falta de datos de entrenamiento. Los autores de

este trabajo proponen un proceso de ajuste fino en dos etapas para la adaptación del dominio. En la primera etapa, el autor afina un modelo usando 17,000 imágenes de orejas de 205 sujetos. Todas las imágenes entre los individuos presentan las mismas posiciones en ángulo e iluminación. En la segunda etapa, los autores realizan un proceso final de ajuste en el conjunto de datos objetivo, en este caso el conjunto de datos UERC, que contiene imágenes de oreja sin restricciones recopiladas de la web. Los autores presentan sus experimentos aplicados a tres modelos: VGG16, AlexNet y GoogLeNet. Los resultados muestran que el modelo que mejor rendimiento alcanza es VGG16 con un valor de precisión de 54,2% empleando el ajuste fino de una etapa, mientras que el proceso de ajuste fino de dos etapas aumenta la precisión a 63,62%. Concluyen que el proceso de adaptación del dominio es realmente necesario y útil para ajustar un modelo pre-entrenado a la tarea de reconocimiento de oreja.

Si bien algunos investigadores utilizaron técnicas de ajuste fino y aplicación de aumento de datos para ayudar a aumentar la precisión, otros investigadores recurrieron a la extracción de características artesanales como complemento para incrementar las tasas de reconocimiento. Hansley, Segundo, & Sarkar (2018) presentaron una técnica para describir imágenes de orejas usando una fusión de vectores de características que fueron extraídos usando técnicas artesanales y técnicas basadas en aprendizaje profundo. Como primer paso en su proceso, los autores entrenan un detector de puntos de referencia basado en una CNN. La arquitectura de la red recibe como entrada una imagen en escala de grises y genera un vector de 110 valores que representa las coordenadas 2D de 55 puntos de referencia. Luego, utilizando los puntos de referencia detectados, se aplica una normalización geométrica entre todas las imágenes del conjunto de datos. Para obtener las características de una imagen de oreja, por un lado, se obtuvieron vectores de características usando técnicas artesanales como Patrones Binarios Locales o Local Binary Patterns (LBP), Características de Imágenes

Estadísticas Binarizadas o Binarized Statistical Image Features (BSIF), Cuantificación de Fase Local o Local Phase Quantization (LPQ), Cuantificación de Fase Local Invariante a la Rotación o Rotation Invariant Local Phase Quantization (RILPQ), Patrones de Procesamiento de Bordes Orientados o Patterns of Oriented Edge Processing (POEM), Histograma de Gradientes Orientados o Histogram of Oriented Gradients (HOG) y Transformación de Características Invariantes a la Escala o Scale Invariant Feature Transform (SIFT); y por otro lado, se entrenó una CNN desde cero para luego usarla como extractor de características. Posteriormente, usando la técnica de Análisis de Componentes Principales o Principal Component Analysis (PCA), todos los vectores de características fueron reducidos a un mismo tamaño para finalmente fusionar ambos descriptores. Los autores validaron su trabajo usando el conjunto de datos AWE, con un valor de 75.6% Rank-1, y usando el conjunto UERC alcanzaron un valor Rank-1 de 49.06%. En ambos casos el mejor resultado fue alcanzado por la fusión de la CNN y los vectores de características generados usando la técnica de HOG. Los autores concluyen que incluir un paso de normalización y emplear la fusión de vectores de características aumenta el rendimiento del reconocimiento.

De la misma manera, Kacar & Kirci (2019) presentó un modelo llamado ScoreNet, que básicamente trabaja en tres etapas. En la primera etapa, se crea un grupo de modalidades. Una modalidad se entiende como una sola operación en un proceso estándar de reconocimiento, como el redimensionamiento de la imagen de entrada, el preprocesamiento, la representación de la imagen (usando técnicas artesanales o de aprendizaje profundo), la reducción de la dimensionalidad y la medición de distancia. En el siguiente paso, las modalidades se seleccionan de forma aleatoria para cada paso en proceso de reconocimiento, generando una secuencia aleatoria de modalidades de reconocimiento de oreja. Luego, la secuencia se aplica en un conjunto de entrenamiento y validación para

generar una matriz de puntuación de similitud. Una vez que se generan múltiples secuencias de este tipo, un algoritmo llamado Deep Cascade Score Level Fusion (DCSLF), propuesto en este trabajo, selecciona los mejores grupos de modalidades y calcula los pesos de fusión necesarios. En el último paso, todos los grupos de modalidad seleccionados (procesos) y los pesos de fusión calculados en la estructura de red en cascada se fijan para el conjunto de datos de prueba. Los autores informan una métrica de rendimiento Rank1 de 61,5% en el conjunto de datos de la UERC. Los autores también muestran que la implementación de procedimientos de normalización y orientación basada en puntos de referencia aumenta el rendimiento general. No obstante, esta arquitectura necesita de un largo entrenamiento y una gran diversidad en el grupo de modalidades. Estos últimos enfoques se presentaron en la competencia de Unconstrained Ear Recognition Challenge 2019 (Emersic, y otros, 2019), siendo las dos técnicas principales que utilizaron enfoques híbridos (técnicas de aprendizaje profundo y técnicas tradicionales) y obtuvieron los mejores resultados en la tarea de reconocimiento de orejas.

2.3 Técnicas y Herramientas

2.3.1 Transferencia de Aprendizaje

Entrenar modelos utilizando técnicas DL a menudo requiere grandes cantidades de datos, recursos informáticos (GPU, TPU) y un tiempo de entrenamiento bastante largo, en pocas palabras es costoso. Un modo de simplificar o acortar este proceso es utilizando la técnica de Transferencia de Aprendizaje o *Transfer Learning* (TL). Este método usa el conocimiento obtenido al momento de resolver una tarea y lo aplica para resolver una tarea similar. Debido a la estructura de las redes neuronales, generalmente el primer conjunto de capas contiene funciones de nivel inferior, mientras que las últimas capas contienen funciones de nivel superior. Estas últimas están más relacionadas con el dominio de la tarea en cuestión. La transferencia de aprendizaje implica reutilizar las capas de nivel inferior en

un nuevo problema o dominio para reducir significativamente el tiempo de entrenamiento, datos y recursos informáticos. Por ejemplo, si ya se tiene un modelo que reconoce automóviles, este modelo puede ser usado también para reconocer camiones, motocicletas y otros tipos de vehículos. (Microsoft Documentation, 2022)

2.3.1.1 Afinación, Ajuste Fino

Un proceso de afinación o *fine tune* entrena una red previamente entrenada por más iteraciones en un conjunto de datos objetivo. De esta forma, se pretende adaptar las representaciones ya aprendidas sobre un conjunto de datos inicial a una tarea diferente. Por lo general, en un proceso de ajuste fino, el conjunto final de capas densas se elimina y se reemplaza con un nuevo conjunto de capas que se ajusta a las clases en el conjunto de datos de objetivo. Entonces podemos mantener algunos pesos fijos y solo entrenar las nuevas capas, o podemos entrenar la red completa hasta la convergencia. (Roman, 2020)

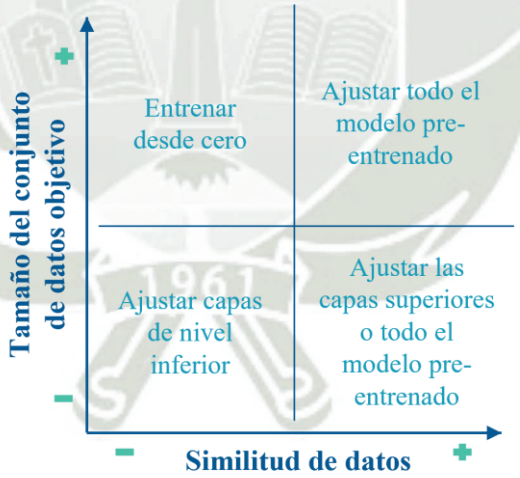

Figura 15 Posibles escenarios en un proceso de afinación de modelo

Fuente: Elaboración propia

Escenario 1: Cuando el tamaño de los datos es pequeño y la similitud de los datos es muy baja. En este caso, podemos congelar las capas iniciales del modelo pre-entrenado y entrenar solo las capas restantes por completo. Así, las capas superiores se personalizarían con el nuevo conjunto de datos objetivo.

Escenario 2: Cuando el tamaño del conjunto de datos es pequeño y la similitud de los datos es muy alta. En este caso, dado que la similitud de los datos es muy alta, no es necesario volver a entrenar el modelo. Todo lo que tenemos que hacer es personalizar y modificar las capas de salida según las clases en el conjunto objetivo. Usamos el modelo pre-entrenado como un extractor de características. Las capas densas y la capa de clasificación son reemplazadas con capas según las nuevas categorías a predecir.

Escenario 3: Cuando el tamaño del conjunto de datos es grande, sin embargo, la similitud de los datos es muy baja. En este caso, entrenar desde cero la red neuronal resultará más efectivo ya que se tiene una gran cantidad de datos.

Escenario 4: Cuando el tamaño de los datos es grande y hay una gran similitud de datos. En este caso, el modelo pre-entrenado debería ser más efectivo que entrenar desde cero. La mejor manera de usar el modelo es retener la arquitectura del modelo y los pesos iniciales del modelo. Luego entrenar este modelo mismo utilizando los pesos inicializados en el modelo pre-entrenado.

2.3.2 Marcos de Trabajo

2.3.2.1 TensorFlow

Es una librería de código abierto desarrollada por el grupo de investigación Google Brain y es activamente usada en Google tanto para investigación como para sus aplicaciones en producción. Fue liberada como código abierto en el 2015. En el año 2019, se lanzó la segunda versión, TensorFlow2.0, una actualización importante que simplificó la biblioteca y la hizo más fácil de usar, además de incluir Keras como parte de su librería. Keras es una biblioteca de Redes Neuronales de código abierto escrita en Python. TensorFlow cuenta con una gran cantidad de herramientas empleadas en el campo de aprendizaje automático. Así como también. para el desarrollo móvil, cuenta con un API para JavaScript y Swift, lo que

le permite comprimir y optimizar modelos para dispositivos móviles e Internet de las cosas (Abadi, y otros, 2015).

Muchos conjuntos de datos y algoritmos de aprendizaje automático populares están integrados en TensorFlow y están listos para usar. Además de los conjuntos de datos integrados, puede acceder a los conjuntos de datos de Google Research o usar la búsqueda de conjuntos de datos de Google para encontrar aún más.

Un inconveniente es que la actualización de TensorFlow 1.x a TensorFlow 2.0 cambió tantas funciones que actualizar el código resulta tedioso y propenso a errores.

2.3.2.2 Jupyter

Es un entorno de trabajo que soporta más de 40 lenguajes de programación incluyendo Python y R, lenguajes bastante usados en las áreas de ciencia de datos y computación científica. Jupyter Notebook permite la exploración de datos en un ambiente en el cual se ejecuta código, se observa que sucede y se puede modificar y repetir de forma iterativa. Los usuarios pueden ingresar código o texto natural dándole un aspecto de bloc de notas que puede ser compartido con otros usuarios o una audiencia específica (Project Jupyter, 2022).

2.4 Consideraciones Finales

Este capítulo presentó algunas definiciones generales relacionadas a la inteligencia artificial, el aprendizaje automático, el aprendizaje profundo y las redes neuronales. Las técnicas de redes neuronales han sido ampliamente usadas para resolver problemas de visión computacional como la clasificación de imágenes, la segmentación y la detección de objetos. Además se revisaron algunos trabajos presentados en la literatura para abordar el problema de reconocimiento de orejas en ambientes no controlados. La oreja humana es un rasgo biométrico con ciertas características que la hacen apropiada para ser usada como fuente de información en sistemas biométricos. Finalmente, una de las técnicas ampliamente usadas en el campo del aprendizaje profundo para resolver nuevas tareas en distintos dominios es la transferencia de aprendizaje. Este método permite reutilizar características aprendidas de distintas fuentes de datos para resolver problemas similares.

CAPÍTULO III

3 METODOLOGÍA

Este trabajo tiene como objetivo identificar personas usando imágenes de orejas tomadas en ambientes no controlados. Para lograr este objetivo, se propone reunir diferentes técnicas para cada etapa del proceso de reconocimiento. Este capítulo describe el proceso de reconocimiento en general y luego se describen las técnicas que fueron empleadas en cada etapa.

3.1 Descripción General del Proceso de Reconocimiento

El proceso inicia con la entrada de una imagen tomada en un ambiente no controlado, esta imagen puede ser de cualquier tamaño y resolución. Como primer paso, se detecta la oreja y se segmenta en un cuadro delimitador al borde de la oreja. La imagen de la oreja segmentada se redimensiona para actuar como entrada en la siguiente etapa. En la etapa de extracción de características, la imagen pasa por un modelo CNN previamente entrenado específicamente para la tarea de reconocimiento de orejas, dicho modelo generará un vector descriptor obtenido de las últimas capas del modelo. Finalmente, dicho vector de características que pertenece a una imagen de prueba se compara con los vectores de características de las imágenes almacenadas en el sistema usando la distancia Coseno entre ambos vectores. En este punto, las distancias más cercanas indican que las imágenes pertenecen al mismo personaje, mientras que las medidas de mayor distancia indican que las dos imágenes comparadas pertenecen a individuos diferentes. Un diagrama del proceso de reconocimiento propuesto se muestra en la Figura 16.

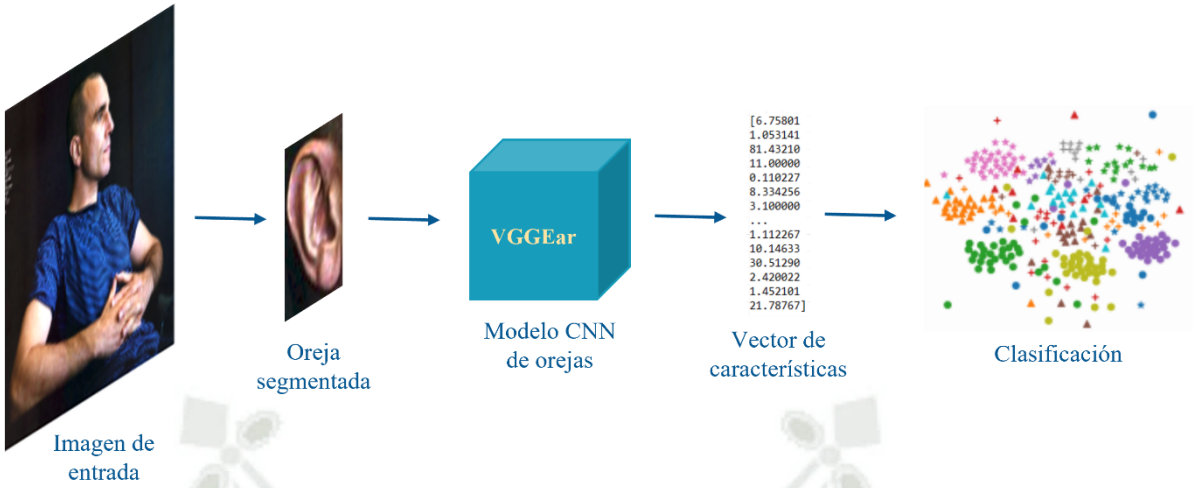

Figura 16 Proceso de reconocimiento de oreja

Fuente: Elaboración propia

3.2 Detección y Segmentación

La primera etapa por la que pasa una imagen es la detección. Dado que el proceso de detección de objetos, en general, ha sido un área de investigación activa durante muchos años, ha alcanzado un gran nivel de madurez. Algunos enfoques han sido descritos en la sección 2.2.2. Enfocándonos en técnicas DL, existe una técnica basada en una CNN llamada Mask R-CNN que tiene muy buen rendimiento en la detección de objetos en ambientes no controlados y una de las técnicas que trata muy bien la oclusión y la sobreposición de objetos.

Para realizar la detección de orejas se usó una implementación disponible en línea de la arquitectura Mask R-CNN (Abdulla, 2019) y tomando como referencia los experimentos presentados por Bizjak, Peer, & Emersic (2019) se implementó la detección de orejas usando Python y Tensorflow. Para el entrenamiento se usó el conjunto de datos de AWE para segmentación, los detalles de este conjunto de datos se encuentran en la sección 2.2.1.2. Se emplearon técnicas de Transferencia de Aprendizaje y Ajuste Fino para adaptar los modelos pre entrenados con grandes conjuntos de datos como ImageNet y CoCo al dominio

de detección de orejas. Se hicieron 4 tipo de experimentos, se utilizaron modelos pre entrenados en ambos conjuntos de datos (ImageNet y CoCo) y por cada conjunto de datos, se probó ajustar todas las capas del modelo o solo las capas superiores o cabeceras. En cuanto a las imágenes de entrada, antes de ingresar a los modelos, todas eran redimensionadas a 1024 x 1024 píxeles conservando la relación de aspecto, rellenando los espacios faltantes con pixeles negros. No se aplicó ninguna técnica de aumento de datos. En cuanto al entrenamiento, la tasa de aprendizaje fue modificada de acuerdo con los experimentos en la implementación.

Una vez que se completa el proceso de detección de orejas, el resultado es una oreja segmentada a nivel de pixeles y el cuadro delimitador de la oreja, elementos que serán usados como entrada para la siguiente etapa.

3.3 Extracción de Características

La segunda etapa en el reconocimiento de personas es la extracción de características. Una técnica ampliamente empleada en el reconocimiento de imágenes en conjuntos de datos desconocidos o abiertos es entrenar una CNN con una gran cantidad de datos tal que luego pueda ser usada como extractor de características y no como un algoritmo de clasificación. En otras palabras, cuando se entrena un modelo para identificar o clasificar un conjunto de datos cerrado o específico, cada vez que se desea agregar un nuevo sujeto al conjunto de datos se tendría que reentrenar el modelo completamente para poder incluir esta nueva identidad. Sin embargo, si se entrena un modelo con un conjunto de datos a gran escala con la suficiente variabilidad para poder generalizar una gran cantidad de características presentes en el conjunto de datos de entrenamiento, no sería necesario hacer un reentrenamiento cada vez que se desee incluir una nueva identidad. Para esto, se usó un modelo que fue entrenado específicamente con este propósito y fue presentado por Ramos-Cooper, Gomez-Nieto, & Camara-Chavez (2022). El modelo, llamado VGGEar, fue

ajustado al dominio de orejas usando un modelo pre-entrenado de rostros. Sin embargo, antes del entrenamiento del modelo en sí, se hicieron tres experimentos previos para determinar la imagen de entrada para la red. Los elementos de salida de la anterior etapa fueron una oreja segmentada a nivel de píxel y un cuadro delimitador de la oreja, ambos podrían ser usados como entrada para el modelo de entrenamiento. No obstante, el tamaño de la imagen de entrada para la CNN debe ser cuadrada, por lo tanto era necesario evaluar diferentes formas de trasformar las imágenes para que puedan ajustarse al tamaño de entrada del modelo. La Figura 17 muestra todas las posibles transformaciones que podría sufrir una imagen de entrada, mientras que las imágenes en Figura 18 muestran las tres imágenes que fueron usadas para entrenar un modelo y analizar cuál sería la mejor opción de segmentación y preprocesamiento.

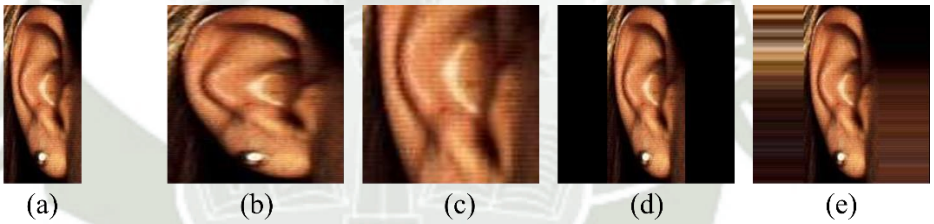

Figura 17 Preprocesamiento de imagen para adaptarla al tamaño de entrada de los modelos CNN.

La primera imagen de la izquierda es la imagen original, las demás son imágenes preprocesadas: (b) redimensionada (c) recortada (d) relleno con pixeles negros (e) relleno con pixeles extendidos del borde la imagen.

Fuente: Elaboración propia

Figura 18 Diferentes formas de segmentación para la oreja

Fuente: Elaboración propia

3.4 Reconocimiento de Personas

El reconocimiento de personas fue dado según la distancia calculada entre dos vectores de características de dos imágenes distintas. El modelo descriptor de orejas presentaba tres capas densas de tamaños 4096, 4096 y 2622, las cuales podrían ser usadas como vectores de características. Sin embargo, de acuerdo con experimentos realizados, se observó que la capa central de tamaño 4096 resultaba en una mejor tasa de acierto. La explicación que se podría dar a esto es que la primera capa densa, la que se encuentra después de las capas convolucionales contiene características que no son tan útiles para la discriminación entre diferentes clases o sujetos. Por otro lado, la tercera capa densa de tamaño 2622 es la capa previa a la capa de clasificación del modelo, esta capa contiene características más específicas para hacer la diferenciar las clases presentes en el conjunto de datos de entrenamiento, en este caso son características ajustadas a las personas con la cuales se entrenó el modelo. Por lo tanto la capa central resulta en un punto central o neutral que hace posible usar las características generadas hasta esa capa para poder describir imágenes de sujetos que nunca fueron "vistas" por el modelo.

Este vector obtenido del modelo CNN es comparado con otros vectores en una base de datos para poder definir la similitud entre los vectores. La técnica usada para esto fue la similitud coseno, la cual está definida por $1 - distancia\ coseno$. A mayor similitud entre dos imágenes, la distancia coseno es menor, mientras que a menor similitud, la distancia es mayor. Una mayor similitud entre dos imágenes se podría interpretar como verificación del mismo sujeto.

3.5 Sistema de reconocimiento

Para poder hacer un completa evaluación de todas las técnicas empleadas en los procesos de detección, segmentación y extracción de características, simulamos el proceso completo

de un sistema de reconocimiento biométrico basado en orejas. La Figura 19 muestra las etapas presentes en un proceso de reconocimiento general.

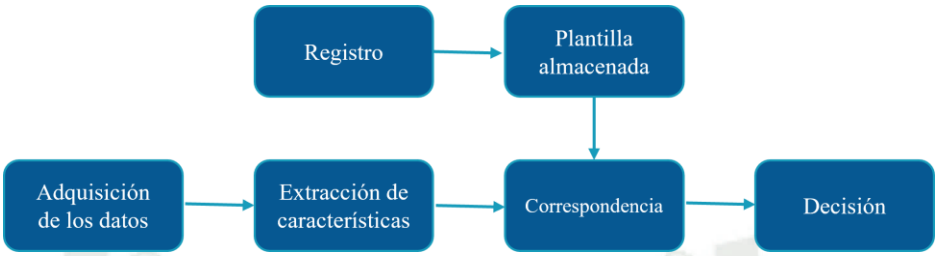

Figura 19 Etapas en un proceso de reconocimiento

Fuente: Elaboración propia

3.5.1 Registro

Para el registro de personas se empleó un método de captura de imágenes desde videos. Se recopilaron videos de 4 personas en distintas ubicaciones (2 interiores, 1 exterior) con diferentes condiciones de iluminación. Se grabaron 8 videos por persona. De los cuales se eligieron 500 imágenes representativas de orejas, para ser parte de la base de datos del sistema. La Figura 20 muestra un diagrama de cómo fueron las sesiones de captura de los videos.

3.5.2 Plantilla almacenada

Las 500 imágenes elegidas para cada persona fueron usadas para extraer características, obteniendo 500 vectores de características que serán almacenados como plantilla para poder identificar a los sujetos.

3.5.3 Adquisición de los datos

Para los experimentos en este trabajo se recopilaron imágenes de la galería de fotos personal del autor, las mismas 4 personas que fueron parte del proceso de registro. La descripción de este conjunto de datos se encuentra en la sección Bases de Datos 4.1. Estas imágenes pasaron

por el detector de orejas implementado y descrito en la sección 3.2. Luego las imágenes detectadas y segmentadas fueron descritas usando el extractor de características descrito en 3.3.

3.5.4 Correspondencia

Para verificar la identidad de una nueva imagen de prueba, se usó la distancia coseno entre los vectores de características de la imagen de prueba con las imágenes almacenadas en la base de datos del sistema.

3.5.5 Decisión

Finalmente la decisión o predicción de la identidad de la persona presente en la imagen de prueba está dada por la identidad de la persona a la cual el vector de características es el más cercano o de mayor similitud.

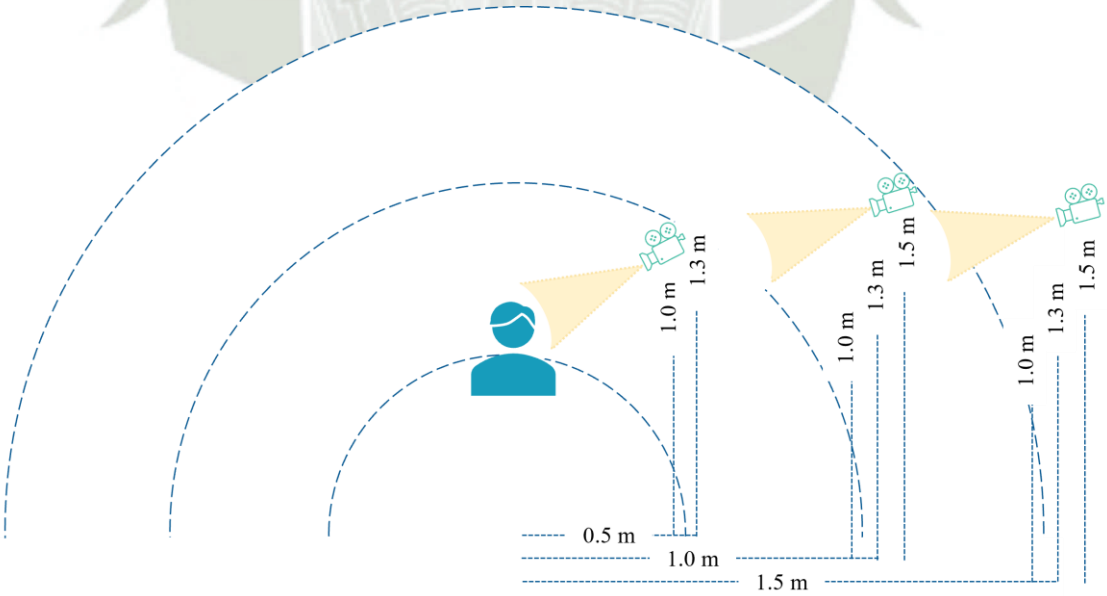

Figura 20 Captura de videos para el proceso de registro de personas

Fuente: Elaboración propia

3.6 Consideraciones Finales

Para poder implementar un sistema de reconocimiento de personas basado en imágenes de orejas se analizaron e implementaron diferentes modelos para ser usados en los procesos de detección, segmentación y extracción de características. En este capítulo se describió los métodos y técnicas empleados para dichos propósitos. En el siguiente capítulo se evaluará el rendimiento de cada modelo, así como el rendimiento de todos los modelos en un proceso end-to-end de reconocimiento.

CAPÍTULO IV

4 PRUEBAS Y ANÁLISIS DE RESULTADOS

4.1 Bases de Datos

A continuación se describe los conjuntos de datos que fueron usados para entrenar y probar cada una de las etapas del proceso de reconocimiento. Cabe destacar que los conjuntos que fueron utilizados en los experimentos, todos ellos recolectan imágenes tomadas en ambientes no controlados, sin tener especial cuidado sobre la pose, iluminación o tamaño de la imagen.

Para el proceso de detección y segmentación de imágenes, se usó el conjunto de datos AWE para segmentación (ver sección 2.2.1.2). Este conjunto fue usado para entrenar y evaluar el modelo Mask-RCNN. Este conjunto de datos contine 1,000 imágenes, de las cuales 900 fueron usadas para entrenar el modelo y 100 fueron usadas para evaluar. Es necesario mencionar que, los conjuntos de entrenamiento y prueba difieren de los conjuntos propuestos por los autores del conjunto de datos, esto se hizo con la sencilla razón de usar más ejemplos para poder ajustar aún más el modelo e incrementar la tasa de detección en cualquier tipo de imágenes de diferentes conjuntos de datos, además que el modelo Mask-RCNN es un modelo que fue probado con altas tasas de acierto para la detección de más de 1,000 objetos diferentes. También, cabe recordar que el proceso de detección es crucial para hacer un correcto reconocimiento, lo cual se verá en la evaluación completa del proceso de reconocimiento.

Para la evaluación del proceso completo de reconocimiento (desde la detección hasta el reconocimiento), se recolectaron imágenes de personajes peruanos públicos y cercanos al

autor. Este conjunto agrupa imágenes de 16 personas con un promedio de 30 ejemplos por persona. Algunas muestras encontradas en este conjunto son mostradas en la Figura 21.

Figura 21 Conjunto de imágenes recolectado por el autor

Fuente: Elaboración propia

4.2 Métricas de Evaluación

El rendimiento de los algoritmos de detección y reconocimiento serán analizados y evaluados usando las siguientes métricas cuantitativas:

a) IoU (*Intersection over Union* o Intersección sobre Unión) es una métrica usada para medir la exactitud de un detector de objetos o cualquier algoritmo que proporcione cuadros delimitadores o *bounding boxes* como salida. Para esto se necesita la información del cuadro delimitador real y el cuadro delimitador proporcionado por el algoritmo. Para hacer el cálculo de esta métrica se debe obtener la razón entre el área de sobreposición y el área de unión de ambos cuadros delimitadores. La Figura 22 muestra

un ejemplo de cuadros delimitadores y la Figura 23 las áreas que son consideradas en el cálculo.

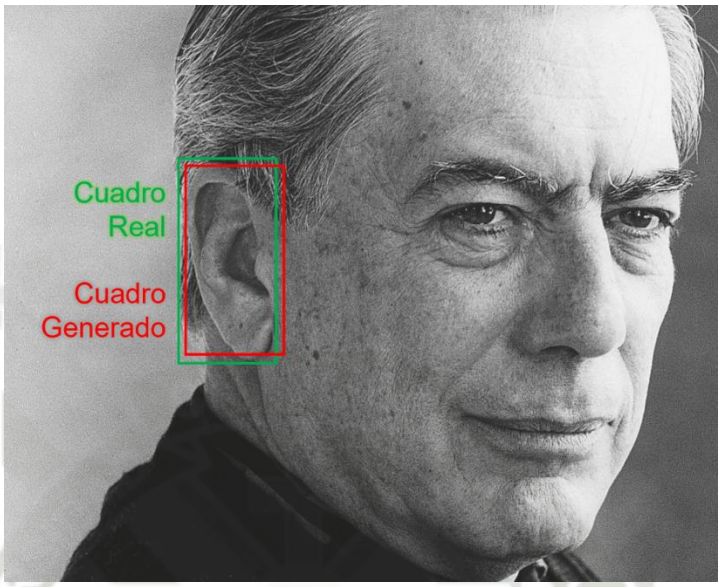

Figura 22 Ejemplo de detección de oreja. El cuadro delimitador en verde y el cuadro delimitador generado en rojo. El objetivo es calcular la Intersección sobre la Unión entre estos cuadros delimitadores.

Fuente: Elaboración propia.

$$IoU = \frac{\text{Área de superposición}}{\text{Área de unión}}$$

Figura 23 Calcular la intersección sobre la unión dividiendo el área de superposición entre el área de unión de los cuadros delimitadores.

Fuente: Elaboración propia.

b) Precisión o *Precision* que expresa el porcentaje de positivos que realmente se han detectado como positivos. La exhaustividad o *Recall* que expresa el porcentaje de elementos positivos que el modelo ha sido capaz de detectar. La exactitud o *Accuracy* que brinda el porcentaje de casos en los que se ha acertado la detección.

$$Precisión = \frac{TP}{(TP + FP)}$$

$$Exhaustividad = \frac{TP}{(TP + FN)}$$

$$Exactitud = \frac{TP + TN}{(TP + FP + TN + FN)}$$

Donde:

- TP (*True Positive* o Verdaderos Positivos) representa los ejemplos correctamente detectados como positivos.
- FP (*False Positive* o Falsos Positivos) representan los ejemplos incorrectamente detectados como positivos.
- TN (*True Negative* o Verdaderos Negativo) representa los ejemplos correctamente detectados como negativos.
- FN (*False Negative* o Falsos Negativos) representa los ejemplos incorrectamente detectados como negativos.

Es importante mencionar que estas métricas pueden también ser ajustadas a una detección a nivel de píxel. Donde: los Verdaderos Positivos o *TP* representan el número de píxeles que pertenecen al objeto y que fueron correctamente detectados como parte del objeto. Los Falsos Positivos o *FP* representan el número de píxeles que pertenecen al objeto, pero no fueron correctamente detectados. Los Verdaderos Negativos o *TN* representan el número de pixeles que no pertenecen al objeto y que fueron correctamente detectados como no parte del objeto. Finalmente, los Falsos Negativos o *FN*, representan la cantidad de pixeles que no pertenecen al objeto y que fueron detectados como parte del objeto.

c) *Rank-1* y *Rank-5* son métricas usadas para medir la tasa de reconocimiento. *Rank-1* representa el porcentaje de imágenes de prueba para las cuales la identidad correcta fue encontrada en la primera coincidencia de las imágenes obtenidas desde la galería,

mientras que *Rank-5* representa el porcentaje de imágenes de prueba para las cuales la identidad correcta fue encontrada entre las 5 primeras coincidencias obtenidas de la galería.

4.3 Experimentos y Resultados

La implementación de los modelos fue hecha usando Python 3.6 y Tensorflow 2.0. El entramiento de los modelos se realizó usando un computador de escritorio con una GPU NVIDIA GeForce RTX 2060. Las características de los datos, modelos e hiper parámetros son detallados a continuación.

4.3.1 Detección y Segmentación de Orejas

Para la detección de orejas se entrenó el modelo Mask-RCNN usando el conjunto de datos de AWE. Se probó entrenar solo las capas superiores o caberas y todas las capas de la arquitectura principal, modelo ResNet101. El modelo implementado se encontraba pre-entrenado con los conjuntos de datos CoCo (Lin, y otros, 2015) e ImageNet (Stanford Vision Lab, Stanford University, 2020). Para ajustar el modelo pre-entrenado y hacer posible la detección de orejas, se tomaron 900 imágenes aleatorias del conjunto AWE para entrenar el modelo, y el resto de las imágenes fueron usadas para probar la precisión del modelo. Se hizo un primer ajuste de 100 épocas y luego un segundo ajuste de 200 épocas más del modelo que mejor rendimiento tuvo en el primer ajuste. Para el entrenamiento se usó el optimizador SGD (Stochastic Gradient Descendent) y con una tasa de aprendizaje de 0.001. Los resultados fueron evaluados usando el cuadro delimitador y se muestran en las Tabla *1* y 2. La métrica usada para la evaluación de este modelo es la precisión, donde los elementos correctamente detectados, o verdaderos positivos, son aquellos en los que la relación de la IoU es mayor o igual a 0.75, y los objetos detectados como orejas pero que no son realmente orejas, son considerados como los falsos positivos. En la Tabla *1* se puede observar que la

técnica que mejor resultados obtuvo fue el de ajustar todas las capas del modelo pre-entrenado con el conjunto de datos CoCo. Se probó ajustar aún más este modelo por 200 épocas adicionales y los resultados obtenidos mejoraron de manera ligera (Ver Tabla 2).

Tabla 1 Ajuste de Modelo Mask-RCNN

		Precisión [%]	
Capas ajustadas	**Conjunto de preentrenamiento**	**Entrenamiento**	**Validación**
Cabeceras	CoCo	97.75	87.50
	ImageNet	90.97	77.25
Todas	CoCo	**98.81**	**91.00**
	ImageNet	93.89	84.25

Fuente: Elaboración propia

Tabla 2 Ajuste de Mask-RCNN por 300 épocas

		Precisión Promedio [%]	
Capas ajustadas	**Conjunto de preentrenamiento**	**Entrenamiento**	**Validación**
Todas	CoCo	**99.50**	**92.50**

Fuente: Elaboración propia

En cuanto al tema de segmentación, se hicieron experimentos para evaluar cual sería la mejor opción para la imagen de entrada al modelo CNN en la etapa de reconocimiento. Se evaluaron tres opciones, la primera presenta la oreja recortada a nivel de píxel, sólo la región

de la oreja mientras que el resto de los pixeles son pintados de color negro. Esta opción no incluye información de contexto en absoluto (Ver

Figura 24). La segunda opción presenta la oreja recortada con un cuadro ajustado que delimita por completo la oreja. Esta opción incluye poca o muy poca información del contexto tal como color de cabello, accesorios de oreja o de cabeza. Por último, la tercera opción presenta la región cuadrada completa de la oreja, obtenida desde la imagen original. En esta opción incluye mucha o bastante información de contexto, según la rotación de la cabeza. En este caso, por ejemplo, si el rostro de la persona se encuentra de frente habrá poca información de oreja y bastante información de contexto que podría incluir otros rostros, orejas u otros objetos. Para el entrenamiento se utilizó un parte de los datos del conjunto VGGFace-Ear. Se eligió solo un subconjunto de 60 sujetos para hacer las pruebas. Las clases de este subconjunto contenían entre 254 y 512 ejemplos para cada clase. Se tomaron 60 ejemplos de forma aleatoria para ser parte del conjunto de validación y 60, para ser parte del conjunto de prueba. El resto de los ejemplos de cada clase fueron aumentados a 300 ejemplos para que sean parte del conjunto de entrenamiento. El modelo utilizado fue una VGG16 con pesos pre-entrenados del conjunto ImageNet. A este modelo, se le reemplazaron las dos últimas capas por una capa completamente conectada de 2048 unidades y una segunda capa de 60 unidades con activación softmax, que sería la capa de predicción. Todos los modelos fueron entrenados por 80 épocas con una tasa de aprendizaje de 0.0001 y con el optimizador SGD.

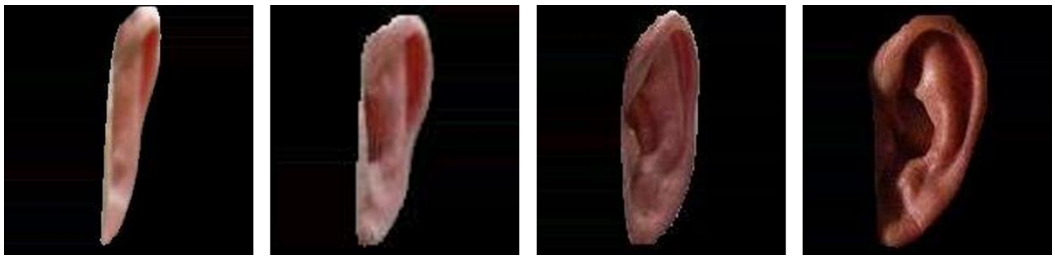

Figura 24 Imágenes de muestra en diferentes ángulos de un mismo sujeto

Fuente: Elaboración propia

La Tabla 3 muestra los experimentos que se hicieron con estas tres opciones. Los resultados muestran que hay un mejor rendimiento usando las opciones 2 y 3, mientras que la opción 1 resulta ser la peor de las tres. Aunque los resultados no son claros en cuanto a cuál sería la mejor opción para ajustar un modelo, se podría entender que quitarle completamente la información de contexto a la imagen de entrada, resulta perjudicial para el reconocimiento. Mientras que conservar poca o mucha información de contexto ayuda en la generalización del modelo para reconocer nuevos ejemplos.

Tabla 3 Resultados de entrenamiento con diferentes imágenes de entrada

	Precisión [%]		**Rank [%]**	
	Validación	**Prueba**	**R1**	**R5**
	90.61	92.22	92.94	97.31
	91.78	**94.00**	**94.86**	98.14
	**93.06**	93.81	94.72	**98.28**

Fuente: Elaboración propia

Para los experimentos de reconocimiento se empleó la segunda opción, una segmentación ajustada al borde de las orejas y con una técnica de preprocesamiento de relleno con pixeles más cercanos al borde. Este proceso fue usado en todas las imágenes, tanto en las de entrenamiento como evaluación.

4.3.2 Reconocimiento de Orejas

Para el reconocimiento de orejas se usó el modelo VGGEar presentado por Ramos-Cooper, Gomez-Nieto, & Camara-Chavez (2022) como extractor de características. Se usó el modelo para obtener los vectores generados en la penúltima capa del modelo, los cuales son de tamaño 4096. Estos vectores luego fueron usados para medir distancias entre las imágenes. Se hicieron pruebas con diferentes técnicas para obtener distancias entre vectores tales como la distancia euclidiana, la distancia coseno y la distancia chi-cuadrado; sin embargo, la distancia coseno proporcionaba mejores resultados. Las imágenes de orejas son representadas se representadas como vectores y la medida entre dos vectores indica la similitud entre las imágenes. En este punto, las distancias pequeñas deberían indicar que las imágenes pertenecen al mismo individuo, mientras que las distancias grandes indicarían que las imágenes pertenecen a diferentes personas. Se hicieron pruebas en el conjunto de imágenes que fue creado específicamente para este trabajo. La Tabla 4 muestra los resultados para 3 conjuntos diferentes. Set1 hace referencia al conjunto que recolecta imágenes de personajes públicos peruanos y que fueron descargadas desde la web, este conjunto consta de 12 sujetos, y con 30 imágenes por sujeto. El set2 junta imágenes que fueron tomadas desde la galería personal del autor. Todas las imágenes fueron tomadas usando celulares móviles, de distintas gamas y calidades. Este conjunto consta de 4 personas y también recolecta 30 imágenes por persona. Finalmente el set3, es la unión de los dos conjuntos anteriores. Es necesario resaltar que hizo esta discriminación en los conjuntos ya que existe un sesgo entre las imágenes recolectadas desde la web con las imágenes tomadas con celulares móviles. Las imágenes del primer conjunto en su mayoría son de alta resolución, de muy buena calidad, probablemente tomadas con cámaras profesionales y buen enfoque. Sin embargo, a pesar de haber dicha diferencia entre las imágenes de los dos

conjuntos, el extractor de características muestra buen rendimiento en ambos conjuntos por separado y juntos.

Para obtener los valores mostrados en la Tabla 4 se siguió el siguiente procedimiento:

a) Se extrajo los vectores de características de todas las imágenes usando el modelo VGGEar.
b) Se tomó el vector de características de una imagen, imagen de prueba, y esta fue comparada con el resto de las imágenes del conjunto, imágenes de la galería.
c) La clase o "sujeto" a la cual pertenece la imagen de galería más cercana a la imagen de prueba es la clase de predicción para la imagen de prueba.
d) En el caso de la métrica Rank, se tomó el primer y los 3 primeros sujetos más cercanos a la imagen de prueba.
e) En el caso de la métrica de precisión, se usó el algoritmo de los *K* vecinos más cercanos para obtener la clase de predicción, donde se tomó *K=1,3,5*.

Tabla 4 Evaluación del proceso de reconocimiento usando métricas Rank y Precisión.

	Rank [%]		**Precisión [%]**		
	R1	**R3**	**K = 1**	**K = 3**	**K = 5**
Set1	96.33	99.67	96.33	97.33	97.00
Set2	98.00	100.00	98.00	98.00	96.00
Set1-2	95.00	99.00	95.00	95.75	94.75

Fuente: Elaboración propia

4.3.3 Evaluación end-to-end del proceso de reconocimiento

Para hacer una evaluación de un proceso end-to-end de reconocimiento fue necesario incluir el proceso de registro de personas en una base de datos que posteriormente se consultaría

teniendo la información de la oreja detectada de un nuevo sujeto. Para esto se hizo el registro de las personas pertenecientes al conjunto Set2, se obtuvo un conjunto de imágenes representativas y los vectores descriptores de estas imágenes serían usadas como plantillas de consulta para la identificación de un nuevo sujeto. Las etapas de un proceso de reconocimiento fueron detalladas en la sección 3.5.

En la etapa de registro, se obtuvieron imágenes muestra de los sujetos mediante sesiones de video, de estos videos, 500 imágenes fueron elegidas para ser parte de la base de datos de consulta luego de extraer los vectores de características de dichas imágenes.

El proceso de reconocimiento inicia con la adquisición de los datos dada una imagen de entrada (50 fotografías de la galería personal del autor). Para esto se usó el modelo detector de orejas obteniendo una tasa de acierto de 85.15%. Se hizo un conteo manual de los objetos que deberían ser detectados como orejas en relación con la cantidad de objetos detectados correctamente. En las 50 fotografías usadas de prueba se encontraban 101 orejas, de las cuales se detectaron correctamente 86 de un total de 94 objetos detectados.

Seguidamente se usó el extractor de características para obtener los vectores que serán comparados con las plantillas almacenadas. Para el proceso de correspondencia o emparejamiento se usó, una vez más, la distancia coseno entre vectores. La Tabla 5 Tasa de reconocimiento para el proceso completo de reconocimiento muestra las tasas de acierto medidas en Rank-n y Precisión. Es importante mencionar que en las fotografías de prueba hay sujetos que no se encuentran registrados, por ello se agregó una clase más que debería abarcar a los sujetos impostores o desconocidos detectados en las imágenes. La Figura 25 muestra la matriz de confusión con la cantidad de aciertos y desaciertos para cada una de las 5 clases registradas en el sistema.

Tabla 5 Tasa de reconocimiento para el proceso completo de reconocimiento

	Rank [%]		Precisión [%]		
	R1	**R3**	**K = 1**	**K = 3**	**K = 5**
Set2	79.17	82.29	79.17	78.12	79.17

Fuente: Elaboración propia

Finalmente, para hacer un análisis visual de la correcta correspondencia de los sujetos de prueba con los sujetos registrados en el sistema, se graficaron los vectores de características de todas las imágenes en un espacio bidimensional (Ver Figura 26). Las etiquetas 101-104 y 999 pertenecen a las 500 imágenes seleccionadas y almacenadas como plantillas de consulta, mientras que las etiquetas 1101-1104 y 1999 son los objetos detectados en las 50 fotografías de prueba.

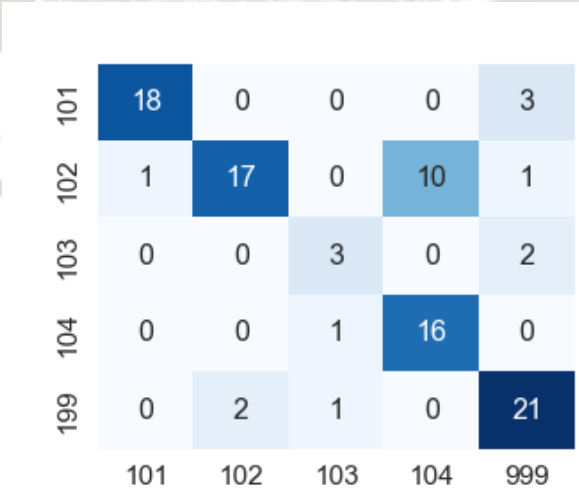

Figura 25 Matriz de confusión para las clases en el conjunto "Set2"

Fuente: Elaboración propia

4.4 Discusión

El rendimiento de cada tarea realizada dentro de un proceso de reconocimiento tiene un impacto en la tasa de acierto del resultado final. En este capítulo se ha descrito la validación de los modelos que fueron entrenados para las tareas específicas de detección y reconocimiento de imágenes de orejas tomadas en ambientes no controlados.

En el caso de la detección, esta tarea fue realizada usando un modelo pre-entrenado en imágenes generales (cosas, animales, automóviles). Si bien existen muchos modelos de detección pre-entrenados con diferentes conjuntos de datos en la literatura, este modelo, Mask-RCNN, fue elegido pensando en la siguiente etapa del proceso de reconocimiento, la segmentación. El modelo Mask-RCNN tiene la capacidad de detectar y segmentar a nivel de píxel un objeto dentro de una imagen. La segmentación a nivel de píxel de una oreja fue inicialmente una propuesta que permitiría eliminar información no deseada o poco útil para la etapa de reconocimiento, como cabello, accesorios, partes de rostro, etc. Sin embargo, después de realizar los experimentos se pudo comprobar de información de contexto es de echo necesaria para que el modelo de reconocimiento pueda tener un buen rendimiento. Por lo tanto, retirar completamente la información de contexto a una imagen de oreja, resulta perjudicial, mientras que conservar poca o mucha información de contexto ayuda en la generalización del modelo de reconocimiento.

En el caso de la etapa de reconocimiento, una técnica que ha sido bastante explorada y con buenos resultados es usar un modelo CNN como extractor de características para obtener un vector descriptor que pueda ser comparado con vectores de diferentes imágenes. Este tipo de extractores de características ha resultado ser más eficiente en imágenes sin restricciones en comparación con técnicas manuales de extracción, tales como Patrones Binarios Locales (LBP), Cuantificación de Fase Local (LPQ), Patrones de Procesamiento de Bordes Orientados (POEM) e Histograma de Gradientes Orientados (HOG), entre otros. Usar un

modelo pre-entrenado y/o adaptado para reconocimiento de orejas, como un extractor de características tiene la ventaja de generar un sistema escalable, ya que no es necesario re-entrenar toda la red para poder realizar el reconocimiento de una nueva identidad. Esta característica asemeja estos modelos a las técnicas de extracción manuales pero con la robustes de un modelo de aprendizaje profundo.

Finalmente, en un proceso end-to-end existen muchos otros factores que podrían deteriorar la tasa de reconocimiento final. Primero, cuando el modelo de detección detecta imágenes que no son orejas, el modelo de reconocimiento siempre dará un resultado erróneo. Segundo, si la segmentación no es realizada apropiadamente y se incluye demasiada información de contexto que no es conocida por el modelo de entrenamiento, esto también causa un deterioro en la tasa de reconocimiento. Tercero, la calidad y enfoque de las imágenes también tienen un impacto significativo en ambas tareas, detección y reconocimiento. Las tablas Tabla 4 y Tabla 5 muestran el deterioro en la tasa de acierto de una evaluación de solo reconocimiento con imágenes de orejas ya recortadas, con la evaluación completa del proceso de reconocimiento que implica la detección, segmentación y reconocimiento en imágenes tomadas en ambientes no controlados, esto incluye contextos variables, diferentes distancias desde la cámara, diferentes tipo de iluminación, diferentes identidades (conocidas y desconocidas), posible oclusión (total o parcial), etc. El deterioro es de alrededor el 20% en un conjunto que incluye 5 identidades diferentes.

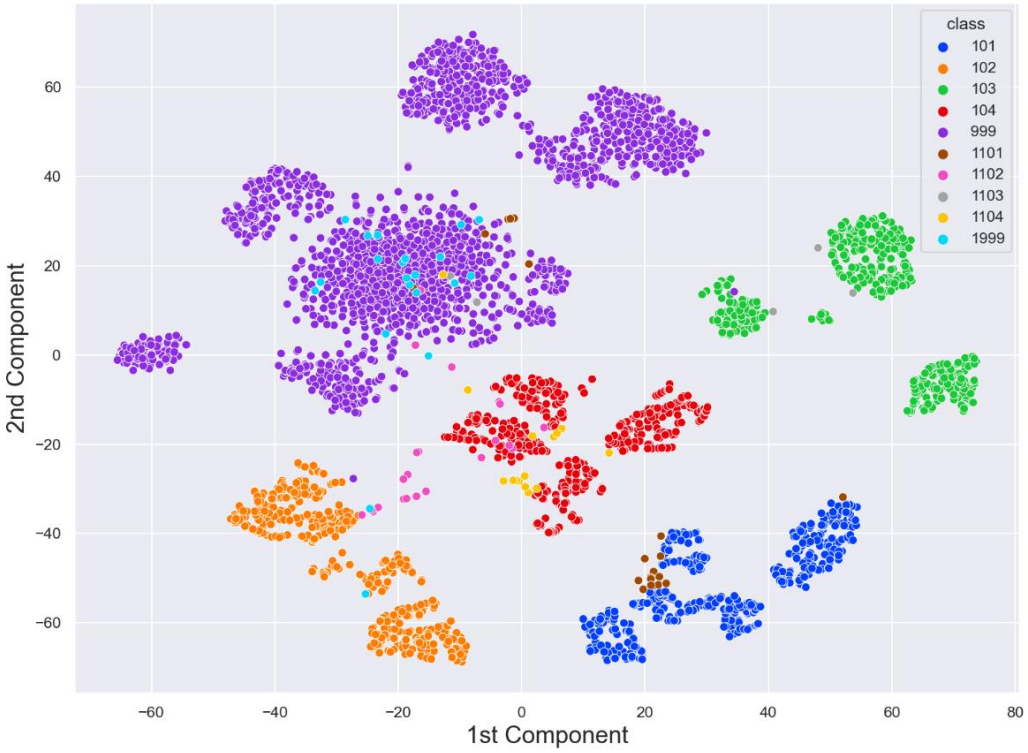

Figura 26 Visualización 2D de los vectores de características

Fuente: Elaboración propia

CONCLUSIONES

i. El reconocimiento de personas usando imágenes de orejas es una tarea que viene siendo ampliamente investigada en los últimos años. Por otro lado, las técnicas de inteligencia artificial y de redes neuronales han demostrado ser efectivas en tareas como detección, segmentación y reconocimiento de imágenes tomadas en ambientes no controlados. La aplicación de este tipo de técnicas en el campo del reconocimiento de orejas es un área de investigación activa y que trae muchas expectativas en el campo de la biometría y en aplicaciones de seguridad.

ii. Según la revisión de distintos trabajos encontrados en la literatura referidos al tema de reconocimiento de orejas, hasta hace algunos años el estado del arte en este campo estaba siendo liderado por diferentes técnicas manuales de visión por computador que obtenía buenos resultados en imágenes tomadas en ambientes controlados. No obstante, al aplicar estas técnicas en imágenes más realistas o tomadas en ambientes exteriores o sin control, resultan en un deterioro en las tasas de acierto.

iii. Una de las técnicas usadas para el entrenamiento de redes neuronales que evita el entrenamiento desde cero, es el aprendizaje por transferencia. Cuando se aplican técnicas de aprendizaje profundo es necesario contar con el modelo adecuado para la tarea y con una cantidad considerable de datos de entrenamiento. Sin embargo, cuando no se tienen grandes cantidades de datos, reusar características aprendidas en modelos pre entrenados resulta ser una opción con buenos resultados. En este trabajo se usó métodos de transferencia de aprendizaje para las tareas de detección y extracción de características usando modelos pre entrenados en conjuntos de datos como ImageNet, CoCo y VGGFace.

iv. Los modelos entrenados y presentados en este trabajo son efectivos. En el caso de la tarea de detección de orejas, el modelo alcanza una tasa de acierto del 92.50% en un conjunto de validación de 100 imágenes. Mientras que en el caso de la tarea de reconocimiento de orejas el modelo entrenado alcanza el 95% de tasa de acierto en un conjunto de 480 imágenes. En ambos casos las imágenes pertenecen a conjuntos de datos que recolectan imágenes tomadas en ambientes no controlados.

v. Evaluar un proceso end-to-end para el reconocimiento de personas usando imágenes de orejas permite saber el impacto que cada proceso tiene sobre la tasa de reconocimiento final. En el caso de la detección y segmentación de orejas, usar un cuadro delimitador con un método de preprocesamiento que no deforme geométricamente la imagen de oreja resultan como mejor opción para esta primera etapa. La extracción de características usando un modelo pre entrenado en un conjunto de datos a gran escala realiza una mejor descripción de la imagen que permite la correcta discriminación de las imágenes por sujeto.

RECOMENDACIONES Y TRABAJOS FUTUROS

i. El tema de alineación y normalización de imágenes previo a la extracción de características es una tarea que debe ser abordada en profundidad para saber si logra tener un impacto positivo y en que intensidad, en la tasa final de reconocimiento. Algunos trabajos en la literatura han abordado este problema, sin embargo existen muy pocos trabajos que lo hayan hecho empleando imágenes tomadas en ambientes no controlados

ii. Establecer un correcto protocolo para el registro de las imágenes que serán usadas como plantillas para posteriores consultas, también debe ser evaluado y definido en un entorno más realista y con conjuntos de datos más grandes. Así como también, aplicar técnicas de reducción de dimensionalidad posteriores a la extracción de características debería ser evaluado y analizado también.

iii. Finalmente, hacer análisis y pruebas en grandes cantidades de datos podría permitir definir un valor umbral que sea considerado para evaluar si dos imágenes pertenecen o no a la misma persona. De esta manera el sistema se podría escalar al reconocimiento de un conjunto de personas mayor y no a un grupo finito de personas registradas en un sistema.

REFERENCIAS BIBLIOGRÁFICAS

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Zheng, X. (2015). *Tensorflow*. Obtenido de Large-Scale Machine Learning on Heterogeneous Systems: https://www.tensorflow.org/

Abdulla, W. (31 de 03 de 2019). *Mask RCNN*. Obtenido de https://github.com/matterport/Mask_RCNN

Aguilar, L. A., & Gallego, A. M. (2013). *Realidad Aumentada en el Museo.* Buenos Aires (Argentina): Universidad Nacional de La Plata.

Akcay, S., Kundegorski, M., Willcocks, C., & Breckon, T. (2018). Using Deep Convolutional Neural Network Architectures for Object Classification and Detection Within X-Ray Baggage Security Imagery. *IEEE Transactions on Information Forensics and Security*, 2203-2215. doi:10.1109/TIFS.2018.2812196

Alcarria Izquierdo, C. (2010). *Desarrollo de un sistema de Realidad Aumentada en dispositivos móviles.* Valencia.

Alshazly, H., Linse, C., Barth, E., & Martinetz, T. (2020). Deep convolutional neural networks for unconstrained ear recognition. *IEEE Access*, 170295-170310.

Ansari, S., & Gupta, P. (2007). Localization of Ear Using Outer Helix Curve of the Ear. *2007 International Conference on Computing: Theory and Applications (ICCTA'07)*.

Attarchi, S., Faez, K., & Rafiei, A. (2008). A New Segmentation Approach for Ear Recognition. *Advanced Concepts for Intelligent Vision Systems*, (págs. 1030-1037). Berlin.

Azuma, R. (1997). A survey of augmented reality. *Teleoperators and Virtual Environments*, 355-385.

Bertillon, A. (1890). *La photographie judiciaire: avec un appendice sur la classification et l'identification anthropométriques.* Paris: Gauthier-Villars.

Bizjak, M., Peer, P., & Emersic, Z. (2019). Mask R-CNN for Ear Detection. *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, (págs. 1624-1628).

Booysens, A., & Viriri, S. (2022). Ear Biometrics Using Deep Learning: A Survey. *Applied Computational Intelligence and Soft Computing*, 1-17.

Booysens, A., & Viriri, S. (2022). Ear biometrics using deep learning: A survey . *Applied Computational Intelligence and Soft Computing*.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

CANDELA. (2016). *CANDELA*. Obtenido de Consortium for the Analysis of the Diversity and Evolution of Latin America: https://www.ucl.ac.uk/candela

Cao, Q., Shen, L., Xie, W., Parkhi, O., & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. *International Conference on Automatic Face and Gesture Recognition.*

Chang, K.-E., Chang, C.-T., Hou, H.-T., Sung, Y.-T., Chao, H.-L., & Lee, C.-M. (2014). Development and behavioral pattern analysis of a mobile guide system with augmented reality for painting appreciation instruction in an art museum. *Computers & Education*, 185-197.

Chen, C.-Y., Chang, B., & Huang, P.-S. (2013). Multimedia augmented reality information system. *Pers Ubiquit Comput*, 315–322.

Chidananda, P., Srinivas, P., Manikantan, K., & Ramachandran, S. (2015). Entropy-cum-Hough-transform-based ear detection using ellipsoid particle swarm optimization. *Machine Vision and Applications*, 185-203.

Chung, N., Han, H., & Joun, Y. (2015). Tourists' intention to visit a destination: The role of augmented reality (AR) application for a heritage site. *Computers in Human Behavior*, 588–599.

Cianciarulo, D. (2015). From local traditions to "augmented reality". The MUVIG Museum of Viggiano (Italy). *Procedia - Social and Behavioral Sciences*, 138-143.

Cintas, C., Delrieux, C., Navarro, P., Quinto-Sánchez, M., Pazos, B., & Gonzalez-Jose, R. (2019). Automatic Ear Detection and Segmentation over Partially Occluded Profile Face Images. *Journal of Computer Science and Technology*.

Cintas, C., Quinto-Sánchez, M., Acuña, V., Paschetta, C., De Azevedo, S., Silva de Cerqueira, C., . . . Delrieux, C. (2017). Automatic ear detection and feature extraction using Geometric Morphometrics and convolutional neural networks. *IET Biometrics*, 211-223.

Computer Vision Laboratory, University of Ljubljana. (2015). *Ear Recognition Research*. Obtenido de http://awe.fri.uni-lj.si/

Deepak, R., Vasant, A., & Manikantan, K. (2016). Ear detection using active contour model. *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, (págs. 1-7). Pudukkottai.

Elsevier Inc. (2015). *Augmented Reality.* Elsevier B.V.

Emersic, Z., Gabriel, L., Struc, V., & Peer, P. (2018). Convolutional encoder--decoder networks for pixel-wise ear detection and segmentation. *IET Biometrics*, 175-184.

Emersic, Z., Gabriel, L., Struc, V., & Peer, P. (2018). Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation. *IET Biometrics*.

Emersic, Z., Krizaj, J., Struc, V., & Peer, P. (2019). Deep Ear Recognition Pipeline. En *Recent Advances in Computer Vision* (págs. 333-362).

Emersic, Z., Kumar, A., Harish, B., Gutfeter, W., Khiarak, J., Pacut, A., . . . Struc, V. (2019). The Unconstrained Ear Recognition Challenge 2019. *2019 International Conference on Biometrics (ICB 2019)*.

Emersic, Z., Meden, B., Peer, P., & Struc, V. (2018). Evaluation and analysis of ear recognition models: performance, complexity and resource requirements. *Neural computing and applications*, 1-16.

Emersic, Z., Stepec, D., Struc, V., & Peer, P. (2017). Training Convolutional Neural Networks with Limited Training Data for Ear Recognition in the Wild. *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, (págs. 987--994).

Emeršič, Ž., Štruc, V., & Peer, P. (2017). Ear recognition: More than a survey. *Neurocomputing*, 26-39.

Eyiokur, F., Yaman, D., & Ekenel, H. (2018). Domain adaptation for ear recognition using deep convolutional neural networks. *IET Biometrics*, 199--206.

Facebook's AI Research lab (FAIR). (2016). *PyTorch*. Obtenido de https://pytorch.org/

Frejlichowski, D., & Tyszkiewicz, N. (2010). The West Pomeranian University of Technology Ear Database -- A Tool for Testing Biometric Algorithms. *Proceedings of International Conference Image Analysis and Recognition*, (págs. 227-234). Povoa de Varzim, Portugal.

Gallego, A., & Aguilar, L. (2013). Realidad Aumentada en el Museo. *Tesina de Licenciatura*. Argentina.

Ganapathi, I. I., Prakash, S., Dave, I., & Bakshi, S. (2020). Unconstrained ear detection using ensemble-based convolutional neural network model. *Concurrency Computation*, 5197.

Girshick, R. (2015). Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*, (págs. 1440-1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, (págs. 580-587).

Gonzalez, E., Alvarez, L., & Mazorra, L. (2019). *AMI Ear Database*. Obtenido de https://ctim.ulpgc.es/research_works/ami_ear_database/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

Google Brain Team. (2015). *TensorFlow*. Obtenido de www.tensorflow.org

Halawani, A., & Li, H. (2016). Human Ear Localization: A Template-Based Approach. *International Journal of Signal Processing Systems*, 258-262.

Hansley, E., Segundo, M., & Sarkar, S. (2018). Employing fusion of learned and handcrafted features for unconstrained ear recognition. *IET Biometrics*, 215--223.

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, (págs. 2980--2988).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, (págs. 2980-2988).

Hopkins, D. (2013). *QR Codes in Education.*

Iannarelli, A. V. (1989). *Ear Identification.*

Instituto Nacional de Estadística e Informática. (15 de Octubre de 2015). *inei.gob.pe*. Obtenido de http://www.inei.gob.pe/estadisticas/indice-tematico/sociales/

Jain, A., Flynn, P., & Ross, A. (2010). *Handbook of Biometrics.* Springer Publishing Company, Incorporated.

Kacar, U., & Kirci, M. (2019). ScoreNet: deep cascade score level fusion for unconstrained ear recognition. *IET Biometrics*, 109--120.

Kamboj, A., Rani, R., & Nigam, A. (2022). A comprehensive survey and deep learning-based approach for human recognition using ear biometric. *Visual Computer, 38*, 2383-2416. doi:10.1007/s00371-021-02119-0

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*.

Kumar, A., & Wu, C. (2012). Automated human identification using ear imaging. *Pattern Recognition*, 956-968.

Kysela, J., & Štorková, P. (2015). Using augmented reality as a medium for teaching history and tourism. *Procedia - Social and Behavioral Sciences*, 926 – 931.

Liang, Y., He, R., Li, Y., & Wang, Z. (2019). Simultaneous segmentation and classification of breast lesions from ultrasound images using Mask R-CNN. *2019 IEEE International Ultrasonics Symposium (IUS)*, 1470-1472. doi:10.1109/ULTSYM.2019.8926185

Lin, T., Maire, M., Belongie, S., Bouderv, L., Girshick, R., Hays, J., . . . Dollar, P. (2015). *COCO Common Objects in Context*. Obtenido de https://cocodataset.org/

Medus, L., Saban, M., Francés-Víllora, J., Bataller-Mompeán, M., & Rosado-Muñoz, A. (2021). Hyperspectral image classification using CNN: Application to industrial food packaging. *Food Control*.

Microsoft Documentation. (2022). *Deep learning vs. machine learning in Azure Machine Learning*. Obtenido de https://docs.microsoft.com/en-us/azure/machine-learning/concept-deep-learning-vs-machine-learning

Milgram, P., & Kishino, F. (1994). Taxonomy of Mixed Reality Visual Displays. *IEICE Transactions on Information and Systems*, 1321-1329.

Mitteroecker, P., & Gunz, P. (2009). Advances in Geometric Morphometrics. *Evolutionary Biology*, 235-247.

MUVIG Museo delle Tradizioni Locali di Viggiano. (15 de 10 de 2015). *MUVIG Museo delle Tradizioni Locali di Viggiano*. Obtenido de www.facebook.com/muvig

Nejati, H., Zhang, L., Sim, T., Martinez-Marroquin, E., & Dong, G. (2012). Wonder ears: Identification of identical twins from ear images. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, (págs. 1201--1204). Tsukuba, Japan.

Pflug, A., & Busch, C. (2012). Ear biometrics: a survey of detection, feature extraction and recognition methods. *IET Biometrics*, 114-129.

Prakash, S., & Gupta, P. (2012). An efficient ear localization technique. *Image Vision Comput.*, 38-50.

Prakash, S., Jayaraman, U., & Gupta, P. (2008). Ear Localization from Side Face Images using Distance Transform and Template Matching. *2008 First Workshops on Image Processing Theory, Tools and Applications*, (págs. 1-8).

Ramos-Cooper, S., Gomez-Nieto, E., & Camara-Chavez, G. (2022). VGGFace-Ear: An Extended Dataset for Unconstrained Ear Recognition. *Sensors*.

Raposo, R., Hoyle, E., Peixinho, A., & Proença, H. (2011). UBEAR: A dataset of ear images captured on-the-move in uncontrolled conditions. *Proceedings of 2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, (págs. 84-90). Paris, France.

Raveane, W., Galdamez, P., & González, A. (2019). Ear Detection and Localization with Convolutional Neural Networks in Natural Images and Videos. *Processes*, 457.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1137-1149.

Rodríguez Fino, E., Martín-Gutiérrez, J., Meneses Fernández, M., & Armas Davara, E. (2013). Interactive Tourist Guide: Connecting Web 2.0, Augmented Reality and QR Codes. *Procedia Computer Science*, 338 – 344.

Roman, V. (27 de 03 de 2020). *towardsdatascience.com/*. Obtenido de https://towardsdatascience.com/cnn-transfer-learning-fine-tuning-9f3e7c5806b2

Ruiz Torres, D. (2011). Realidad Aumentada, Educación y Museos. *Revista de Comunicación y Nuevas Tecnologías, 2*, 212-226.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*.

Stanford Vision Lab. (2015). *ImageNet Large Scale Visual Recognition Challenge*. Obtenido de http://www.image-net.org/challenges/LSVRC/

Stanford Vision Lab, Stanford University. (2020). *ImageNet*. Obtenido de https://www.image-net.org/

Vinh, T. H. (2019). EarVN1.0: A new large-scale ear images dataset in the wild. *Data in Brief*, 2352-3409.

Wang, S., Du, Y., & Huang, Z. (2017). Ear detection using fully convolutional networks. *2nd International Conference on Robotics, Control and Automation*, (págs. 50-55).

Winter, M. (2011). *Scan Me: Everybody's Guide to the Magical World of QR* .

Yuan, L., & Mu, Z.-C. (2007). Ear Detection Based on Skin-Color and Contour Information. *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007*, (págs. 2213-2217).

Zhang, Y., & Mu, Z. (2017). Ear Detection under Uncontrolled Conditions with Multiple Scale Faster Region-Based Convolutional Neural Networks. *Symmetry*.

Zhou, Y., & Zaferiou, S. (2017). Deformable Models of Ears in-the-Wild for Alignment and Recognition. *Proceedings of 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, (págs. 626-633). Washington, DC, USA.

ANEXOS

i. Proyecto de tesis

ii. Artículo publicado: *VGGFace-Ear: An Extended Dataset for Unconstrained Ear Recognition*

Universidad Católica de Santa María

Facultad de Ciencias e Ingenierías Físicas y Formales

Escuela Profesional de Ingeniaría de Sistemas

DETECCIÓN, SEGMENTACIÓN Y LOCALIZACIÓN DE PUNTOS DE REFERENCIA EN IMÁGENES DE OREJAS CAPTURADAS EN AMBIENTES NO CONTROLADOS USANDO REDES NEURONALES CONVOLUCIONALES

Línea: Inteligencia Artificial

Sub-Línea: Inteligencia Artificial Aplicada

Autora: Solange Griselly Ramos Cooper

Arequipa

Setiembre, 2021

1. PLANTEAMIENTO DE LA INVESTIGACIÓN

Planteamiento del Problema

La detección de oreja es una fase importante dentro de las tareas que engloba el reconocimiento de personas usando imágenes de orejas. Estas tareas están siendo ampliamente investigadas en los últimos años debido a su potencial uso en aplicaciones de seguridad, video vigilancia y ciencias forenses (Pflug & Busch, 2012) (Emeršič, Štruc, & Peer, 2017). Los sistemas de reconocimiento más comunes hacen uso de huellas dactilares, imágenes de rostros y escaneo de iris. Sin embargo, el reconocimiento de personas usando imágenes de orejas, tiene importantes y atractivas ventajas sobre otros sistemas basados en diferentes elementos biométricos. Por ejemplo, la oreja no sufre cambios drásticos a medida que la persona envejece y tampoco la estructura es afectada por las expresiones faciales. De acuerdo con (Jain, Flynn, & Ross, 2010) la estructura de la oreja permanece sin cambios entre la edad de 8 y 70 años, después de los 70 años la oreja crece de forma longitudinal, pero sin cambiar su estructura. Por lo tanto, la oreja es una fuente de información estable y confiable para el reconocimiento de personas. Además, las imágenes de orejas pueden ser capturadas a distancia y sin la explicita cooperación de las personas.

Sin embargo, el reconocimiento de oreja también presenta algunos desafíos que deben ser abordados. Uno de ellos es la oclusión, esta puede ser producida por el cabello, audífonos, auriculares, aretes, piercings, entre otros. Así como otros problemas como la iluminación, posición de la cabeza, imágenes borrosas, baja resolución, entre otros. Estos problemas generalmente se presentan cuando la captura de imágenes se realiza en ambientes no controlados, lo que se conoce como reconocimiento en la naturaleza. El reconocimiento en ambientes no controlados hace referencia al hecho de capturar las imágenes en escenarios de la vida real donde existe una alta variación generada por el entorno. Esta variación puede

generar cambios de iluminación, de enfoque, escala, oclusión, entre otros, lo cual termina siendo un gran desafío para las tareas de detección, segmentación y reconocimiento.

Objetivos de la Investigación

General

Modelamiento de redes neuronales convolucionales para la detección, segmentación y localización de puntos de referencia en imágenes de orejas capturadas en ambientes no controlados.

Específicos

a) Caracterizar el estado del arte de la Inteligencia Artificial en tareas específicas de detección y segmentación de objetos, usando redes neuronales convolucionales.

b) Evaluar diferentes técnicas de detección y segmentación de orejas y analizar trabajos relacionados.

c) Analizar y evaluar técnicas de localización de puntos de referencia para la alineación de orejas y revisar trabajos relacionados.

d) Implementar modelos de redes neuronales convolucionales para la detección, segmentación y localización de puntos de referencia en imágenes de orejas.

e) Validar la efectividad de los modelos propuestos.

Preguntas de Investigación

f) ¿Es posible que técnicas de aprendizaje profundo puedan ser aplicadas a la detección y segmentación de imágenes de orejas?

g) ¿Qué tan efectivo es el uso de técnicas de aprendizaje profundo para la localización de puntos de referencia en objetos?

h) ¿Cuál es el estado del arte en el contexto de detección y segmentación de orejas usando técnicas de aprendizaje profundo?

i) ¿Qué ventajas existen en el uso de técnicas aprendizaje profundo comparadas con técnicas tradicionales aplicadas al reconocimiento en ambientes no controlados?

j) ¿Cuál es la efectividad o tasa de acierto del uso de redes neuronales convolucionales en las tareas de detección, segmentación y localización de puntos de referencia en imágenes de orejas?

Línea y Sub-línea de Investigación

Línea: Inteligencia Artificial

Sub-línea: Inteligencia Artificial Aplicada

Palabras Clave

Detección de orejas, segmentación, alineación, redes neuronales convolucionales, imágenes de orejas, reconocimiento en ambientes no controlados.

Solución Propuesta

Justificación e Importancia

La investigación de técnicas de aprendizaje profundo o *Deep Learning* (DL) y Redes Neuronales Convolucionales o *Convolutional Neural Networks* (CNN) son áreas que vienen siendo exploradas hace varios años y su aplicación en distintos problemas y situaciones del día a día. Las CNN han tenido un gran éxito en aplicaciones prácticas y constituyen el estado del arte en muchos problemas de visión computacional, esto se debe al hecho de que han demostrado ser muy eficaces para la clasificación de imágenes a gran escala (Akcay, Kundegorski, Willcocks, y Breckon, 2018) (Liang, He, Li, y Wang, 2019) (Medus, Saban, Francés-Víllora, Bataller-Mompeán, y Rosado-Muñoz, 2021). Además, una de las competencias más conocidas de visión por computador, el ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (Stanford Vision Lab, 2015), evidencia que los algoritmos basados en técnicas de aprendizaje profundo

alcanzan los mejores resultados en tareas como localización, detección y clasificación de objetos.

Descripción de la Solución

Un proceso completo de reconocimiento usualmente implica las tareas de detección, segmentación, alineación, descripción y finalmente el reconocimiento en sí. En tanto, antes del reconocimiento, es necesario abordar las tareas previas de forma efectiva ya que esto tendrá un impacto en las siguientes fases del reconocimiento. En este trabajo se propone utilizar técnicas DL y CNN para abordar el problema de detección, segmentación y localización de puntos referenciales para finalmente incrementar la precisión en la fase final de reconocimiento. La propuesta implica explorar diferentes modelos, entrenarlos y ajustarlos a las tareas mencionadas.

2. FUNDAMENTOS TEÓRICOS

Estado del Arte

Dentro de los trabajos más recientes presentados en el contexto de detección de orejas se puede encontrar un primer grupo métodos basados en técnicas como detección de bordes, transformación de distancia, contornos activos, gráficos e incluso algunos que usan información de textura y color, entre otras técnicas de visión por computadora. Algunas técnicas fueron ampliamente investigadas en (Attarchi, Faez y Rafiei, 2008) (Prakash, Jayaraman y Gupta, 2008) (Prakash y Gupta, 2012) (Chidananda, Srinivas, Manikantan, y Ramachandran, 2015) (Deepak, Vasant, Manikantan, 2016) (Halawani y Li, 2016). No obstante, este tipo de métodos generalmente no resultan lo suficientemente robustos en escenarios realistas (Zhang and Mu, 2017). Esto puede provocar una caída en las tasas de detección al realizar pruebas dentro de una configuración de entorno variable o imágenes tomadas en la naturaleza.

Dado que los enfoques basados en técnicas tradicionales podrían disminuir el rendimiento de la tarea de detección de oreja debido a condiciones variables, las técnicas basadas en algoritmos de aprendizaje por computadora comenzaron a evaluarse. Una primera investigación de este tipo se presentó en (Islam, Bennamoun y Davies, 2008). El método hace uso del algoritmo AdaBoost para seleccionar los mejores clasificadores débiles y luego combinarlos en clasificadores fuertes, y obtener así una cascada de clasificadores que funciona como detector de oreja. Los autores informan una tasa de detección del 100% en 203 imágenes de rostros de perfil de un conjunto de datos de la Universidad de Notre Dame (UND). No obstante, entrenar este tipo de clasificadores toma varios días. En (Abaza, Hebert y Harrison, 2010) se presentó una mejora a esta técnica, que reduce considerablemente la complejidad de la fase de entrenamiento y la habilita para ser operativa en tiempo real. El método organiza las características de Haar en un clasificador AdaBoost en cascada para detectar orejas en varias poses de perfil e imágenes de oclusiones parciales. Los resultados muestran un 88,5% de precisión utilizando la colección F del conjunto de datos de UND.

Con el incremento en la popularidad de las CNN y su probada efectividad en tareas como la clasificación de imágenes y la detección de objetos, estas arquitecturas comenzaron a desempeñar un papel importante en la tarea de detección de orejas también. Un primer método de detección de orejas basado en una CNN es el trabajo presentado en (Wang, Du y Huang, 2017). Los autores ajustaron la red AlexNet para entrenar un clasificador de orejas, recolectando una gran cantidad de muestras positivas y negativas para este entrenamiento. Luego, convirtieron las capas completamente conectadas en capas completamente convolucionales y utilizaron un enfoque de ventana deslizante para construir el detector de orejas. Los autores informaron una tasa de detección de alta precisión; sin embargo, estas tasas varían según la cantidad de muestras positivas que

utilizaron, es decir, cuantas más muestras, mayor precisión pueden alcanzar. Otro método que, en lugar de utilizar un enfoque basado en ventanas deslizantes, utiliza un enfoque basado en regiones es presentado en (Zhang y Mu, 2017). Los autores proponen un método para recortar la región de la oreja a partir de imágenes mediante la combinación del contexto de ubicación de la oreja y las características morfológicas, concluyendo que la información del contexto mejora el rendimiento de detección. Su método hace uso de una CNN basada en regiones de escala múltiple para reconocer tres regiones en una imagen: cabeza, región "pan-ear" y oreja. Los autores informan tasas de precisión utilizando los conjuntos de datos WebEars, UND-J2 y UBEAR, alcanzando una precisión del 98%, 100% y 98,7% para cada conjunto de datos, respectivamente. En (Ganapathi, Prakash, Dave y Bakshi, 2020) se presentó un enfoque diferente que utiliza CNN también. Este enfoque utilizó un sistema de conjunto de tres CNN para detectar orejas en imágenes faciales de perfil 2D, sin estar restringido a una sola cara. El modelo funciona en dos partes, la primera parte de la técnica entrena tres modelos de CNN en un conjunto de datos dado, luego una parte posterior obtiene un promedio ponderado de la salida de los modelos entrenados para detectar regiones auriculares. Los autores concluyen que un modelo de conjunto supera a los modelos basados en una única CNN. La técnica se validó en dos conjuntos de datos: el IIT Indore-Colection A y el AWE, alcanzando el 98,2% y el 99,5% de precisión en cada conjunto de datos, respectivamente. De la misma manera, (Raveane, Galdamez y Gonzáles, 2019) presentó un método que involucra múltiples CNN para identificar la presencia y ubicación de las orejas en una imagen de entrada dada. Aunque no alcanzan una alta velocidad, su método se probó principalmente en datos de video que reflejan escenarios del mundo real y aplicaciones en tiempo real. Para estos dos últimos enfoques, el principal problema que se debe abordar en trabajos futuros es la disminución de las tasas de falsos positivos. Además, es importante tener en cuenta

que todos los enfoques presentados anteriormente realizan una detección de cuadro delimitador alrededor de la oreja, en otras palabras, la salida de estos métodos son regiones rectangulares que contienen orejas y, en algunos casos, estas regiones no encajan bien en la región de la oreja, lo que agrega algunas áreas no deseadas o información innecesaria a la región de interés que puede afectar procesos posteriores como el reconocimiento.

Por otro lado, algunos enfoques que abordan la tarea de detección de orejas de una manera diferente a la generación de regiones de cuadros delimitadores son las técnicas que realizan una localización de la oreja nivel de píxel. Todos los trabajos a continuación mencionados, fueron evaluados utilizando anotaciones de píxeles a través de la métrica de rendimiento Intersección sobre Unión (IoU), que de hecho representa una relación entre la cantidad de píxeles que están presentes en ambas regiones auriculares detectadas y anotadas, y los píxeles que están en unión de las regiones de la oreja detectadas y anotadas. Un trabajo inicial presentado en (Emersic, Gabriel, Struc y Peer, 2018) propuso una red convolucional de codificador-decodificador con una tarea de segmentación semántica de dos clases, donde el objetivo es asignar cada píxel de la imagen a la clase oreja o sin oreja. La técnica se probó en imágenes 2D de entornos no controlados. El modelo alcanza el 55,7% con respecto a la métrica de rendimiento de IoU. Más tarde, un enfoque de detección presentado en (Emersic, Krizaj, Struc y Peer, 2019) como parte de un proceso completo de reconocimiento de orejas basada en CNNs, mostró una CNN basada en un modelo existente llamado RefineNet. Los autores validaron la técnica en 250 imágenes pertenecientes al conjunto de datos AWE y obtuvieron un valor de IoU del 84,8%.

Finalmente, en (Cintas et al., 2017) se evalúa un enfoque diferente que incluye métodos de Morfometría Geométrica (GM), que son ampliamente utilizados en estudios de

organismos biológicos. Propusieron cuantificar la forma de cada espécimen de acuerdo con la ubicación en el espacio de un conjunto de puntos de referencia 2D o 3D que son homólogos entre los individuos. Los autores entrenaron una CNN con una colección de imágenes anotadas en puntos de referencia basadas en GM para detectar 45 puntos de referencia en imágenes del oído externo, luego usaron un proceso de casco convexo o Convex Hull para obtener el conjunto de píxeles correspondiente solo a la estructura de la oreja. Su método fue probado usando un conjunto de datos utilizado originalmente para el reconocimiento facial y tomado en entornos controlados y configuraciones restringidas, y lo comparó con el detector de orejas basado en Haar disponible en la biblioteca OpenCV (Bradski, 2000). Por lo tanto, este método no se probó en ningún conjunto de datos de reconocimiento de la oreja sin restricciones aún.

Bases Teóricas de la Investigación

Reconocimiento de la Oreja

Recientes investigaciones han demostrado que reconocimiento basado en la oreja no es menos efectiva que otros tipos de reconocimiento, tales como rostro o huellas dactilares. De hecho, el desarrollo de diferentes técnicas ha permitido que el reconocimiento de oreja alcance un rendimiento similar al de reconocimiento basado en imágenes de rostros, por ejemplo. Ha sido también experimentalmente probado que las orejas son únicas en cada individuo (Bertillon, 1890). Además, sus características son más estables a lo largo de vida de una persona que comparado con el envejecimiento de rostros o el daño causado en las huellas digitales debido a tareas repetitivas. Esto contribuye a que las orejas sea un elemento relativamente estable para la identificación y verificación de personas. Estas características sumadas al bajo costo que tiene la adquisición y manejo de datos incrementan aún más el interés en este campo.

La Figura 1 muestra la apariencia de la oreja u oído externo, la cual está definida por el hélix, antihelix, trago, antitrago, la concha, las fosas triangular y escafoidea, los pilares del antihelix y el lóbulo.

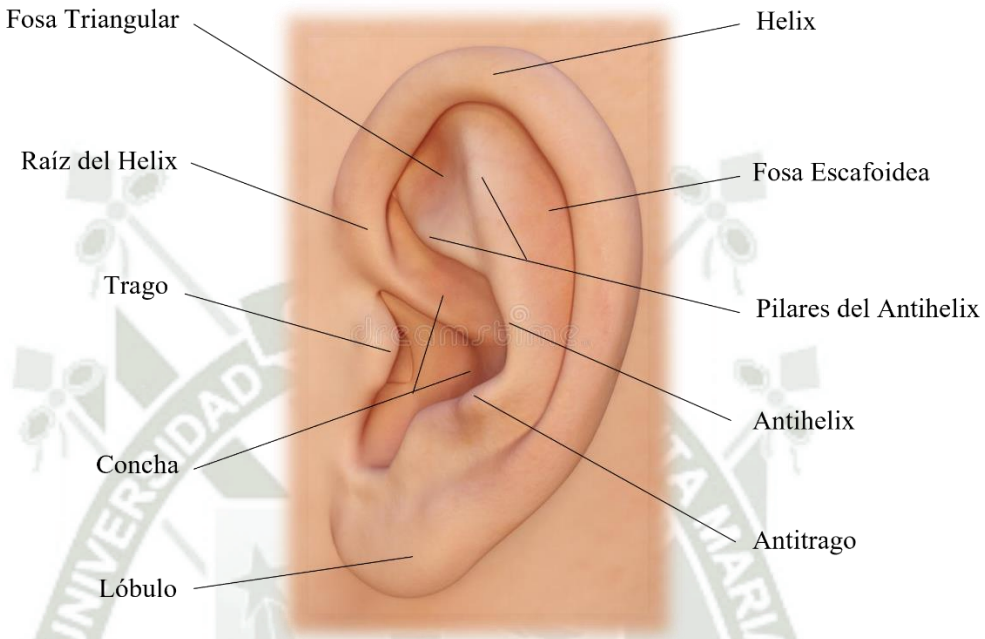

Figura 1 *Estructura de la oreja, imagen elaborada por el autor.*

A pesar de que el reconocimiento de oreja no es usado ampliamente aún para la identificación y/o verificación de personas, existe concreta evidencia que es adecuado y confiable para ser usado e implementado en sistemas biométricos (Pflug y Busch, 2012) (Jain et al., 2010).

Redes Neuronales Convolucionales (CNN)

La red neuronal convolucional, también conocida como red convolucional, es un tipo especializado de red neuronal para el procesamiento de datos que tiene una topología conocida tipo cuadrícula. Un ejemplo claro son los datos de imagen, que pueden pensarse como una cuadrícula 2D de píxeles. El nombre "convolucional" indica que la red emplea una operación matemática llamada convolución que es un tipo especializado de

operación lineal (Goodfellow, Bengio y Courville, 2016). La arquitectura de una CNN típica consiste en dos partes: la parte de extracción de características, que usa operaciones de convolución para extraer características de una imagen; y la parte de clasificación, que es ejecutada haciendo uso de un Perceptrón Multicapa (ver Figura 2).

La parte de clasificación de una CNN se basa en el tradicional modelo de una red neuronal artificial, el cual puede ser definido de la siguiente manera. Dada una red neuronal de N capas, sea X_0 y X_N los vectores de entrada y salida respectivamente de la red. Donde el vector X_{n-1} es la entrada de la capa n (donde n =1, …, N). Además, si W_n es una matriz de pesos y b_n un vector de bias, entonces la capa de salida, X_n, puede ser representada con el siguiente vector:

$$X_n = f\ (W_n\ X_{n-1} +\ b_n),$$

donde f es una función de activación. En un problema de reconocimiento, el proceso de entrenamiento consiste en encontrar los valores óptimos para el conjunto de parámetros de pesos y bias $\{W_n, b_n\}$ que minimice la clasificación de error. Es una práctica común utilizar el algoritmo de gradiente descendiente para determinar cómo deben cambiar los parámetros para disminuir el error. La definición del error depende del tipo de dato de salida (Goodfellow et al., 2016).

La Figura 2 muestra la arquitectura básica de una CNN usada para clasificación. Esta arquitectura presenta 4 tipos de capas principalmente. Capa convolucional, de agrupación o Pooling, unidad lineal rectificada o ReLU, y una capa completamente conectada o Dense Layer. Las capas de convolución toman una imagen como entrada y usan una matriz de filtro para realizar la convolución. La capa de agrupación se usa comúnmente para reducir la dimensionalidad y extraer la información más significativa de cada sección de la imagen. Las capas ReLU se utilizan para llevar la no linealidad a la red.

Finalmente, las capas completamente conectadas se utilizan para conectar todas las neuronas de la capa anterior a la siguiente capa.

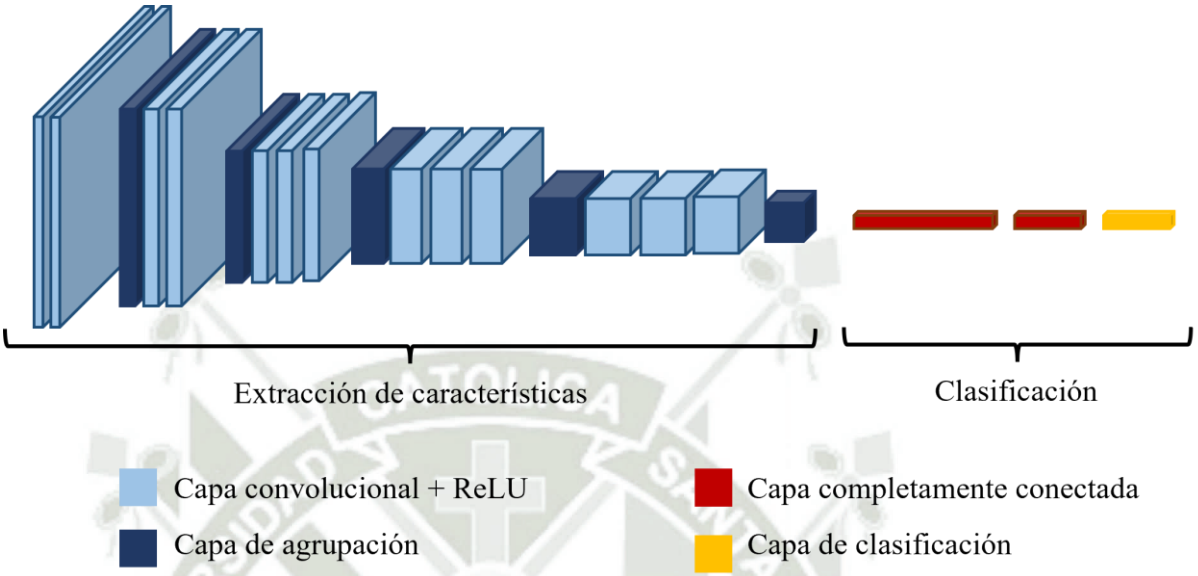

Figura 2 *Estructura de una CNN, imagen elaborada por el autor.*

3. MARCO METODOLÓGICO

Alcances y Limitaciones

En cuanto al alcance de esta investigación, este abarca la evaluación de los principales modelos propuestos en la literatura que hagan uso de técnicas de inteligencia artificial y que hayan sido aplicados en contextos similares como el reconocimiento facial o reconocimiento de otras características biométricas. Mientras que las limitaciones están dadas por la cantidad y disponibilidad de bases de datos propuestas en la literatura para resolver distintas tareas de visión computacional, las cuales son un elemento importante cuando se aplica técnicas de aprendizaje profundo.

Aporte

Los principales aportes de este trabajo son la evaluación de distintas arquitecturas existentes en la literatura aplicadas al contexto de reconocimiento de orejas; y los

modelos entrenados, ajustados y validados para las tareas específicas de detección, segmentación y detección de puntos de referencia en imágenes de orejas.

Tipo y Nivel de la Investigación

El tipo de investigación es aplicada, ya que se pretende investigar el uso de técnicas de aprendizaje profundo y redes neuronales convolucionales para las tareas de detección y segmentación de orejas en imágenes tomadas en ambientes no controlados. Por otro lado, el nivel de investigación es netamente experimental.

Población y Muestra o Universo

La validación de los modelos se llevará a cabo en un muestreo de imágenes que serán recolectadas específicamente para este propósito. Se espera que las imágenes pertenezcan a docentes, administrativos y/o alumnos de la Universidad Católica de Santa María.

Métodos, Técnicas e Instrumentos de Recolección de Datos

Las técnicas de recolección de datos se darán bajo el contexto de reconocimiento en la naturaleza. Esto quiere decir, que las imágenes serán capturadas en espacios al aire libre, sin especial cuidado en temas de iluminación, posición de la cámara respecto a la persona, distancia o temas relacionados al uso de aretes, accesorios de oídos o cabeza u otros elementos.

Plan de Análisis estadístico de los datos

El rendimiento de los algoritmos de detección y segmentación serán analizados y evaluados usando las siguientes métricas cuantitativas:

k) IoU (*Intersection over Union* o Intersección sobre Unión) mide el porcentaje de área que fue detectado correctamente por imagen, de donde se puede obtener un valor promedio para todas las imágenes.

$$IoU = \frac{Área\ de\ superposición}{Área\ de\ unión}$$

l) Precisión que expresa el porcentaje de positivos que realmente se han detectado como positivos. La exhaustividad que expresa el porcentaje de elementos positivos que el modelo ha sido capaz de detectar. La exactitud que brinda el porcentaje de casos en los que se ha acertado la detección.

$$Precisión = \frac{TP}{(TP + FP)}$$

$$Exhaustividad = \frac{TP}{(TP + FN)}$$

$$Exactitud = \frac{TP + TN}{(TP + FP + TN + FN)}$$

Donde:

- TP (*True Positive* o Positivos Verdaderos) representa los ejemplos correctamente detectados como positivos.
- FP (*False Positive* o Falsos Positivo) representan los ejemplos incorrectamente detectados como positivos.
- TN (*True Negative* o Verdadero Negativo) representa los ejemplos correctamente detectados como negativos.
- FN (*False Negative* o Falso Negativo) representa los ejemplos incorrectamente detectados como negativos.

4. PLAN DE TRABAJO

Task	Duración	Comienzo	Fin
Proyecto de investigación sobre detección y segmentación de imágenes de orejas	**96 días**	**vie 1/10/21**	**vie 11/02/22**
Caraterización del estado del Arte	**19 días**	**vie 1/10/21**	**mié 27/10/21**
Lectura y análisis de trabajos relacionados sobre detección y segmentacion de orejas	7 días	vie 1/10/21	lun 11/10/21
Lectura y análisis de trabajos relacionados sobre detección de puntos de referencia en orejas	7 días	mar 12/10/21	mié 20/10/21
Revision y búsqueda de bases de datos disponibles	5 días	jue 21/10/21	mié 27/10/21
Evaluar técnicas existentes	**28 días**	**jue 28/10/21**	**lun 6/12/21**
Evaluación de modelos propuestos en la literatura aplicacidos al contexto de reconocimiento de orejas	14 días	jue 28/10/21	mar 16/11/21
Evaluación de modelos propuestos en la literatura aplicacidos en otros contextos de biometria	14 días	mié 17/11/21	lun 6/12/21
Implementación de modelos	**22 días**	**mar 7/12/21**	**mié 5/01/22**
Implementación de modelos y arquitecturas CNN	7 días	mar 7/12/21	mié 15/12/21
Entrenamiento y ajuste de modelos	15 días	jue 16/12/21	mié 5/01/22
Experimentación y validación	**7 días**	**jue 6/01/22**	**vie 14/01/22**
Evaluación, valiación y comparación de modelos propuestos	7 días	jue 6/01/22	vie 14/01/22
Redacción del Informe de Tesis	**20 días**	**lun 17/01/22**	**vie 11/02/22**
Capitulo de Introducción	5 días	lun 17/01/22	vie 21/01/22
Capítulo de Marco Teorico	5 días	lun 24/01/22	vie 28/01/22
Capítulo de Propuesta	5 días	lun 31/01/22	vie 4/02/22
Capítulos de Experimentos, Discusión y Conclusiones	5 días	lun 7/02/22	vie 11/02/22

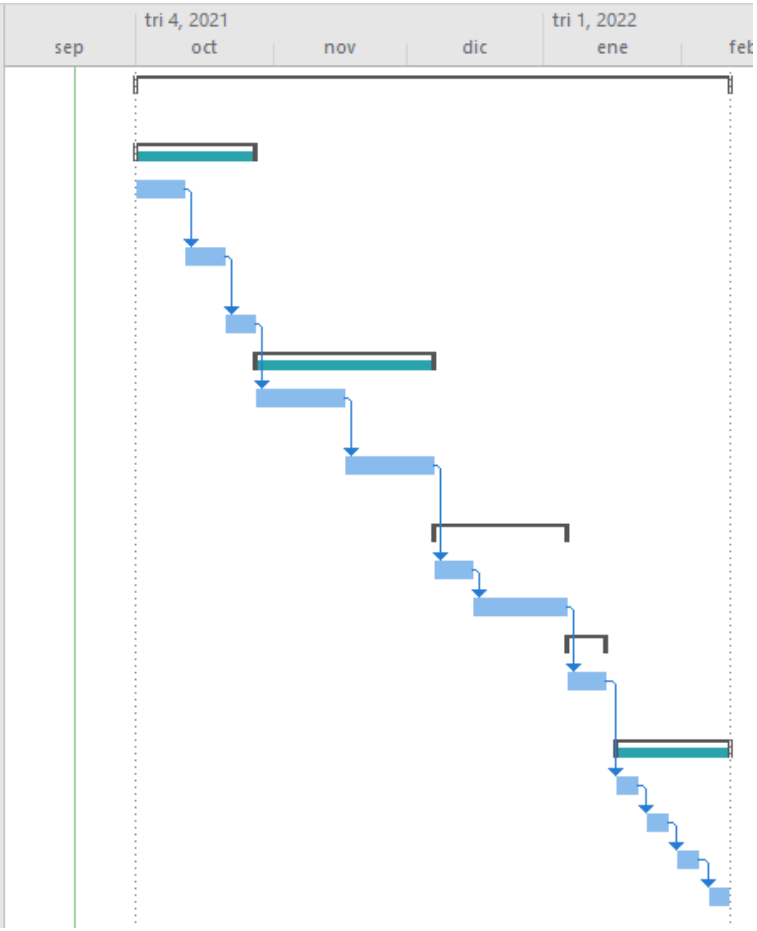

Figura 3 *Diagrama de Gantt para la ejecución del proyecto de Investigación, elaborada por el autor.*

5. REFERENCIAS BIBLIOGRÁFICAS

Abaza, A., Hebert, C., et al. (2010). Fast learning ear detection for real-time surveillance. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6.

Akcay, S., Kundegorski, M. E., et al. (2018). Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. *IEEE Transactions on Information Forensics and Security*, 13(9):2203–2215.

Attarchi, S., Faez, K., et al. (2008). A New Segmentation Approach for Ear Recognition. In Blanc-Talon, J., Bourennane, S., et al., editors, *Advanced Concepts for Intelligent Vision Systems*, pages 1030–1037, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bertillon, A. (1890). *La photographie judiciaire: avec un appendice sur la classification et l'identification anthropométriques*. Gauthier-Villars.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Chidananda, P., Srinivas, P., et al. (2015). Entropy-cum-Hough-transform-based ear detection using ellipsoid particle swarm optimization. *Machine Vision and Applications*, 26:185–203.

Cintas, C., Quinto-Sánchez, M., et al. (2017). Automatic ear detection and feature extraction using Geometric Morphometrics and convolutional neural networks. *IET Biometrics*, 6(3):211–223.

Deepak, R., Vasant, A., et al. (2016). Ear detection using active contour model. Pág. 1–7.

Emersic, Z., Gabriel, L. L., et al. (2018). Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation. *IET Biometrics*, 7(3):175–184.

Emersic, Z., Krizaj, J., et al. (2019). *Deep Ear Recognition Pipeline*, pages 333–362.

Ganapathi, I. I., Prakash, S., et al. (2020). Unconstrained ear detection using ensemble-based convolutional neural network model. *Concurrency Computation*, 32(1):e5197.

Goodfellow, I., Bengio, Y., et al. (2016). *Deep Learning*. The MIT Press.

Halawani, A. and Li, H. (2016). Human Ear Localization: A Template-Based Approach. *International Journal of Signal Processing Systems*, 4:258–262.

Islam, S. M. S., Bennamoun, M., et al. (2008). Fast and Fully Automatic Ear Detection Using Cascaded AdaBoost. In *2008 IEEE Workshop on Applications of Computer Vision*, pages 1–6.

Jain, A. K., Flynn, P., et al. (2010). *Handbook of Biometrics*. Springer Publishing Company, Incorporated.

Liang, Y., He, R., et al. (2019). Simultaneous segmentation and classification of breast lesions from ultrasound images using mask r-cnn. In *2019 IEEE International Ultrasonics Symposium (IUS)*, pág. 1470–1472.

Medus, L. D., Saban, M., et al. (2021). Hyperspectral image classification using cnn: Application to industrial food packaging. *Food Control*, 125:107962.

Pflug, A. and Busch, C. (2012). Ear biometrics: a survey of detection, feature extraction and recognition methods. *IET Biometrics*, 1(2):114–129.

Prakash, S. and Gupta, P. (2012). An efficient ear localization technique. *Image and Vision Computing*, 30:38–50.

Prakash, S., Jayaraman, U., et al. (2008). Ear Localization from Side Face Images using Distance Transform and Template Matching. In *2008 First Workshops on Image Processing Theory, Tools and Applications*, pág. 1–8.

Raveane, W., Galdamez, P., et al. (2019). Ear Detection and Localization with Convolutional Neural Networks in Natural Images and Videos. *Processes*, 7:457.

Stanford Vision Lab (2015). *ImageNet Large Scale Visual Recognition Challenge*. Recuperado de http://www.image-net.org/challenges/LSVRC/

Wang, S., Du, Y., et al. (2017). Ear detection using fully convolutional networks. pages50–55.

Zhang, Y. y Mu, Z. (2017). Ear Detection under Uncontrolled Conditions with Multiple Scale Faster Region-Based Convolutional Neural Networks. *Symmetry*, 9(4).

6. POSIBLE TEMARIO DEL INFORME FINAL

A continuación, el posible índice de contenido del Informe de Tesis.

1. Abstract
2. Resumen
3. Introducción
4. Generalidades
   - 4.1. Identificación del problema
   - 4.2. Objetivo General
   - 4.3. Objetivos Específicos
   - 4.4. Marco Conceptual
     - 4.4.1. Biometría de la Oreja
     - 4.4.2. Redes Neuronales Convolucionales
   - 4.5.Estado del Arte
     - 4.5.1. Detección y Segmentación de Objetos
     - 4.5.2. Detección de Puntos de Referencia
   - 4.6. Métodos y Procedimientos
5. Modelo propuesto
   - 5.1. Visión general de modelo
   - 5.2. Análisis del modelo propuesto
6. Evaluación experimental
7. Análisis y Discusión

sensors

MDPI

Article

VGGFace-Ear: An Extended Dataset for Unconstrained Ear Recognition †

Solange Ramos-Cooper [1], Erick Gomez-Nieto [1] and Guillermo Camara-Chavez [1,2,*]

[1] Department of Computer Science, Universidad Catolica San Pablo, Arequipa 04001, Peru; solange.ramos@ucsp.edu.pe (S.R.-C.); emgomez@ucsp.edu.pe (E.G.-N.)
[2] Computer Science Department, Federal University of Ouro Preto, Ouro Preto 35400-000, Brazil
* Correspondence: guillermo@ufop.edu.br
† This paper is an extended version of our paper published in: Ramos-Cooper, S.; Camara-Chavez, G. Ear Recognition in The Wild with Convolutional Neural Networks. In Proceedings of the 2021 XLVII Latin American Computing Conference (CLEI), Cartago, Costa Rica, 25–29 October 2021.

Abstract: Recognition using ear images has been an active field of research in recent years. Besides faces and fingerprints, ears have a unique structure to identify people and can be captured from a distance, contactless, and without the subject's cooperation. Therefore, it represents an appealing choice for building surveillance, forensic, and security applications. However, many techniques used in those applications—e.g., convolutional neural networks (CNN)—usually demand large-scale datasets for training. This research work introduces a new dataset of ear images taken under uncontrolled conditions that present high inter-class and intra-class variability. We built this dataset using an existing face dataset called the VGGFace, which gathers more than 3.3 million images. in addition, we perform ear recognition using transfer learning with CNN pretrained on image and face recognition. Finally, we performed two experiments on two unconstrained datasets and reported our results using Rank-based metrics.

Keywords: ear recognition; ear biometrics; deep learning; convolutional neural networks; mask-RCNN; VGGFace; transfer learning

Citation: Ramos-Cooper, S.; Gomez-Nieto, E.; Camara-Chavez, G. VGGFace-Ear: An Extended Dataset for Unconstrained Ear Recognition. *Sensors* **2022**, 22, 1752. https://doi.org/10.3390/s22051752

Academic Editors: M. Jamal Deen, Subhas Mukhopadhyay, Yangquan Chen, Simone Morais, Nunzio Cennamo and Junseop Lee

Received: 14 January 2022
Accepted: 14 February 2022
Published: 23 February 2022

1. Introduction

Identifying people is a persistent issue in society. Different areas such as forensic science, surveillance, and security systems usually demand solutions for this issue. Most identification systems implement biometrics to fulfill their requirements. A biometric trait has unique and specific features that make people's recognition possible. Among the most common physical biometric traits such as fingerprints, palmprints, hand geometry, iris, and face, the ear structure results in an excellent source to identify a person without their cooperation. It provides three meaningful benefits, i.e., (i) the outer ear structure does not drastically change over a person's lifetime, (ii) it can be captured from a distance and, (iii) is unique for everyone, even for identical twins [1]. When a person ages, the ear shape does not show significant changes between 8 and 70 years [2]. After the age range of 70–79 years, the longitudinal size slightly increases while the person gets old; however, the structure remains relatively constant. Facial expressions do not affect the ear shape either. the ear image has uniform color distribution and is completely visible even in mask-wearing scenarios. Moreover, the ear structure is less prone to injuries than hands or fingers. These features make the ear structure a stable and reliable source of information used in a biometric system.

The appearance of the outer ear is defined by the lobule, tragus, antitragus, helix, anti-helix, concha, navicular fossa, scapha, and some other critical structural parts, as Figure 1 illustrates. These anatomical characteristics differ from person to person and have also been recognized as a means of personal identification for criminal investigators. In 1890,