DEPARTAMENTO DE
INFORMÁTICA Y AUTOMÁTICA
FACULTAD DE CIENCIAS

VNiVERSiDAD
D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

TESIS DOCTORAL

Sistema biométrico para detección y reconocimiento de orejas basado en algoritmos de procesamiento de imágenes y redes neuronales profundas

AUTOR

D. Pedro Luis Galdámez Canales

DIRECTORA

Dra. Dña. Angélica González Arrieta

Salamanca, España
Junio 2017

La memoria titulada "Sistema biométrico para detección y reconocimiento de orejas basado en algoritmos de procesamiento de imágenes y redes neuronales profundas" que presenta D. Pedro Luis Galdámez Canales para optar al Grado de Doctor por la Universidad de Salamanca ha sido realizada bajo la dirección de la profesora Dra. Dña. Angélica González Arrieta, Profesora Titular del Departamento de Informática y Automática de la Universidad de Salamanca.

Salamanca, junio de 2017

El graduado,

Fdo: D. Pedro Luis Galdámez Canales

La directora,

Fdo: Dra. Dña. Angélica González Arrieta
Profesora Titular de Universidad
Departamento de Informática y Automática
Área de Ciencia de la Computación e Inteligencia Artificial
Universidad de Salamanca

Pedro Luis Galdámez Canales: Sistema biométrico para detección y reconocimiento de orejas basado en algoritmos de procesamiento de imágenes y redes neuronales profundas © *junio 2017*

Supervisora:
Dra. Dña. Angélica González Arrieta

Salamanca, España

Resumen

La oreja es un rasgo biométrico emergente que ha llamado la atención de la comunidad científica por más de una década. Su estructura única ha destacado desde hace mucho tiempo entre los científicos forenses, y se ha utilizado para la identificación de sospechosos en muchos casos. El paso lógico hacia una aplicación más amplia de la biometría de orejas es crear un sistema de reconocimiento. Este trabajo se centra en el uso de datos de imágenes (2D) para llevar a cabo dicho proceso. El presente estudio aborda técnicas como la distancia Hausdorff; la cual agrega robustez e incrementa el desempeño filtrando los sujetos a utilizar en la etapa de prueba del proceso. También incluye la transformación de imágenes con rayos (IRT) en la etapa de detección. La oreja es una característica biométrica inconstante cuando se trabaja con imágenes fotográficas en condiciones variables: esto se debe en gran parte al enfoque de la cámara, las formas irregulares de las capturas, las condiciones de iluminación y la forma siempre cambiante de la proyección cuando es fotografiada. Por tanto, para identificar la presencia y localización de una oreja en una imagen proponemos un sistema de detección de orejas con múltiples redes neuronales convolucionales (CNN) y un algoritmo de agrupación de detección.

El método propuesto coincide con el rendimiento de otras técnicas cuando analizamos fotografías limpias, es decir, capturas en condiciones ideales (purpose-shot), alcanzando una precisión de más del 98 %. Cuando el sistema está sujeto a imágenes naturales en condiciones del mundo real, donde el sujeto aparece en una multitud de orientaciones y condiciones fotográficas en ambiente no controlado, nuestro sistema mantiene la misma precisión superando claramente el resultado del 83 % promedio obtenido en investigaciones previas. Finalmente se exponen los algoritmos utilizados para completar los pasos del reconocimiento, utilizando estructuras convolucionales, técnicas de extracción de características y aproximaciones geométricas a fin de incrementar la presición del proceso.

Abstract

The ear is an emerging biometric feature that has caught the attention of the scientific community for more than a decade. Its unique structure has stood out since long ago among forensic scientists, and has been used to identify suspects in many cases. The logical step towards a broader application of ear biometrics is to create a recognition system. To carry out this process, this work focuses on the use of data from images (2D). The present study mentions techniques like the Hausdorff distance, which adds robustness and increases the performance, filtering the subjects to use in the testing process. It also includes image ray transform (IRT) in the detection step. The ear is a fickle biometric feature when working with photographic images under varying conditions. This is largely due to the camera's focus, the irregular shapes of the captures, the lighting conditions and the ever-changing shape of the projection when it is photographed. Therefore, to identify the presence and location of an ear in an image, we propose an ear detection system with multiple convolutional neural networks (CNN) and a clustering algorithm of detections.

The proposed method coincides with the performance of other techniques when we analyze clean photographs, that is to say, catches in ideal conditions (purpose-shot), reaching an accuracy of more than 98 %. When the system is subjected to natural images in real world conditions, where the subject appears in a multitude of orientations and photographic conditions in an uncontrolled environment, our system maintains the same precision, clearly exceeding the average result (83 %) obtained in previous researches. Finally, the algorithms used to complete the recognition steps are presented, using convolutional structures, extraction techniques and geometric approximations in order to increase the accuracy of the process.

Dedico esta tesis a mi familia y en especial a mi Padre de quien estaré eternamente orgulloso

Agradecimientos

La presente investigación si bien ha requerido de esfuerzo y mucha dedicación, no hubiese sido posible sin el apoyo, aliento y paciencia de todas y cada una de las personas que a continuación citaré, muchas de las cuales han sido un soporte indispensable en el largo camino recorrido.

En primer lugar, me gustaría agradecer a mi directora, por todos los momentos que hemos compartido, por su paciencia, su decisivo apoyo, su guía, por motivarme y levantarme cada vez que lo he necesitado. Gracias Dra. Angélica González por no rendirse conmigo y por mostrarme que siempre hay un próximo día y nuevos retos que superar.

Gracias mil a los compañeros de la Escuela Nacional de Policía, especialmente al Inspector Jefe Miguel Ramón Ramón por su orientación, increíble apoyo y consejo, sin ello esta investigación no habría sido posible. El agradecimiento extendido al equipo de la Comisaría General de Policía Científica que su enseñanza ha sido un baluarte; sin ustedes todo habría sido más largo, complicado y sin duda menos interesante.

Agradecer finalmente hoy y siempre a mi familia, que sin todo su esfuerzo no habría llegado hasta aquí. A mis padres, Carmen y Luis, gracias por creer en mí, a mi abuela Chela por extrañarme y quererme como nadie. A Astrid, por ser mi soporte y mi alegría, tu compañía ilumina mis días, sin ti esto no sería posible.

Tabla de contenido

Listado de figuras

Listado de tablas

Cree a aquellos que buscan la verdad. Duda de los que la encuentran [Croyez ceux qui cherchent la vérité, doutez de ceux qui la trouvent]

André Gide

1 Introducción

Una de las más interesantes características del sistema de visión del ser humano es su habilidad para reconocer rápidamente un estimulo a partir de una gran cantidad de información visual; lo más remarcable es la manera eficiente de cómo este mecanismo trabaja incluso cuando los objetos son tan sólo parcialmente similares a los objetos en memoria. Esta función del cerebro no ha sido completamente comprendida, aunque a veces parece estar claro; el cerebro no compara el estimulo con todos los objetos en la memoria de uno en uno, abstrae la información con el fin de sólo intentar comparar las más significativas características.

En la presente investigación se expone un sistema capaz de llevar a buen termino el reconocimiento de personas a través de imágenes de orejas sin la necesidad de realizar una búsqueda lineal en todos los modelos en la base de datos que representa la memoria visual del sistema. Las clases construidas son capaces de indexar

y comparar lo que permitirá mapear la estructura jerárquica de las partes del objeto en una especie de gráfico, con el propósito exclusivo de intentar simular la abstracción que el cerebro realiza en el proceso de visión e identificación. Como ejemplo de la descomposición de un objeto en partes, se puede considerar la jerárquica relación entre las partes de la silueta del cuerpo de una persona: cuatro largas extremidades (brazos y piernas) y una forma como burbuja (la cabeza) todo subordinado a un tronco. Para descartar las figuras distintas a un cuerpo, la idea es seleccionar solamente aquellos objetos que tengan una configuración similar a la antes descrita para el análisis futuro. Una jerarquía de partes puede también ser dividida por el nivel de detalle que las partes representan, por ejemplo, los dedos se pueden expresar como los detalles de un objeto con forma de mano. Dividir la información en partes nos ayuda a determinar si la persona que lleva un sombrero es aún similar a nuestra abstracción de lo que representa el cuerpo de una persona identificando el sombrero como un accesorio de nuestro modelo.

1.1. Memoria visual

Con el propósito de definir el conjunto de funciones y/o algoritmos de reconocimiento, necesitamos primero decidir la manera más conveniente de representar todas las formas en las que los objetos, orejas, serán probablemente encontrados. Esta aproximación basada en vistas es más simple y no requiere realizar transformación de las imágenes; sin embargo, la base de datos resultante es de gran tamaño, añadiendo a la búsqueda complejidad. Por otro lado los modelos basados en 3D tienen la ventaja además de poseer bases de datos compactas, mantienen información completa soportando vistas no contempladas. Los métodos basados en vistas no requieren una completa descripción de cada modelo; más allá, proporcionan un medio para hacer frente a los objetos con textura, es decir, aquellos objetos cuya estructura no es precisamente plana sino que contienen características como la profundidad, para los que las estadísticas de intensidad de los píxeles también pueden ser considerados.

Un problema en común es la representación requerida para codificar las características relevantes del objeto y sus interrelaciones. Una estrategia en ambos enfoques ha sido descomponer la estructura del objeto en un número pequeño de formas primitivas formando bloques de construcción de todas las vistas posibles. Desafortunadamente, no existe un acuerdo general sobre lo que constituye las partes primitivas de todas las formas, y si hay un conjunto único y pequeño de partes para todos los objetos. Multiples enfoques basados en vistas [54, 55, 170] buscan secciones de características o pequeños parches en la vista del objeto. Estos parches pueden ser aquellos que contienen esquinas, bordes, etc. Sin embargo, esta estrategia no es suficiente para responder la pregunta general de como generalizar las propiedades de un objecto.

1.2. Selección de características

¿Qué características deberíamos utilizar para realizar el proceso de reconocimiento? No hay una respuesta concisa a esta pregunta en la literatura hasta la fecha, donde la respuesta dada dependerá de la clase de objetos que queremos reconocer, y de si queremos identificar casos particulares o cualquier objeto que pertenece a una determinada clase. Es decir, hay algunos objetos que tienen una forma bien definida y un conjunto de partes, pero hay otros que están mejor caracterizados por su textura como por ejemplo el agua y los árboles, los alimentos por su color, los patrones como se ven en libros, estuches, o inclusive su funcionalidad ejemplo sillas, mesas, coches, etc. A su vez, también tenemos que estar de acuerdo en lo que queramos reconocer, por ejemplo, nuestro perro o cualquier perro.

En este estudio, nos hemos centrado en las orejas, las que pueden ser caracterizadas por su forma. Diferentes experimentos psicofísicos apoyan la idea de que un gran número de objetos entran en esta categoría. Trabajos pioneros sobre este tema fueron realizados por Biederman [101, 103], que examina el campo en sus di-

ferentes investigaciones [102]. Como ejemplo de estos tipos de experimentos, podemos considerar el trabajo de Hayward et al. [238], en el que los autores probaron si la gente podía encontrar la información del contorno útil para el reconocimiento de un conjunto de objetos desconocidos desde nuevos puntos de vista. Ellos presentaron a los sujetos una imagen sombreada como estímulo inicial, y luego introdujeron un segundo estímulo de siluetas del objeto. Encontraron que el reconocimiento fue más rápido para imágenes sombreadas sobre las siluetas cuando el objeto estaba en el mismo punto de vista, pero que no había ninguna diferencia en el rendimiento del reconocimiento cuando el objeto giraba con respecto al original. Los resultados sugieren que la información de sombreado no es fundamental para la identificación de los objetos bajo cambios en el angulo de visión.

1.3. Reconocimiento basado en vistas

En la presente investigación, el objetivo es desarrollar un sistema para reconocer de manera eficiente un número potencialmente grande de usuarios independientemente de los cambios de perspectiva en las imágenes. Para ello, vamos a utilizar distintos algoritmos para realizar la detección del objeto. El propósito es obtener una abstracción a fin de descomponer una forma en un conjunto de partes primitivas jerárquicamente organizados.

En nuestros experimentos, en la base de datos de modelos cada una de las vistas se extrae y se utiliza como entrada para la preparación de los algoritmos de detección y reconocimiento. Un grupo de vistas se utiliza para rápidamente realizar una verificación. A continuación, el comparador calcula la similitud entre la consulta y cada uno de los candidatos, y los ordena en consecuencia. El candidato mejor clasificado se considera la vista más similar a la consulta.

1.4. Trabajos relacionados

Los conceptos con respecto a cada uno de los componentes de la investigación, como, el proceso de detección, la representación de los objetos, el algoritmo de coincidencia, el algoritmo de reconocimiento son, en cierta medida, independientes unos de otros. Por lo tanto, preferimos incluir secciones de trabajo relacionados en los capítulos donde se introducen cada uno de los componentes.

Es importante resaltar el trabajo relacionado en su conjunto, es decir, el reconocimiento de objetos de forma general. Varios autores han estudiado este problema utilizando información de sus formas a partir de distintas vistas.

Beis y Lowe [108] realizaron reconocimiento de objetos a partir de la información de su forma utilizando los bordes rectos como los valores primitivos que componen el objeto, agrupándolos de acuerdo a criterios de colinealidad y paralelismo. Los ángulos y proporciones de longitud de estos grupos se utilizan para formar vectores de características que fueron indexados a partir de una base de datos de puntos utilizando árboles KD (k dimensiones); una técnica de búsqueda del vecino más cercano. Los vecinos más cercanos al vector consulta son ponderados de acuerdo a las distribuciones de probabilidad aprendidas en una etapa de entrenamiento, y se utilizan para generar hipótesis. Los modelos que reciban más votos se verifican utilizando un método de ajuste de los modelos 3D parametrizados, descritos por Goldberg [83].

Al igual que nuestro enfoque, Sebastian et al. [223] utilizaron gráficos de choque para representar siluetas de imágenes 2D como muestra de la visualización de un objeto. Los autores, sin embargo, proponen una partición jerárquica de la base de datos, en la que las formas se agrupan en categorías. Un pequeño número de ejemplos de cada categoría son elegidos para representar cada agrupación. Un problema con este enfoque es que no siempre es posible encontrar grupos lo suficientemente grandes como para reducir significativamente el tamaño de la base de

datos. Además, cada coincidencia en el nivel de prototipo implica una búsqueda lineal entre las formas que pertenecen a esa categoría, que debe ser lo más pequeño posible.

En un trabajo reciente, Cyr y Kimia [60] exploraron el problema de cómo dividir la vista de un objeto 3D utilizando una colección de gráficos de choque. Sin embargo, no abordan el problema de indexación, y recurren en cambio a una búsqueda lineal de todas las vistas de la base de datos con el fin de reconocer un objeto.

1.5. Reconocimiento biométrico

Siendo utilizado originalmente para la identificación forense, los sistemas biométricos evolucionaron de ser una herramienta para la investigación criminal a ser una serie de aplicaciones comerciales. Los medios tradicionales de reconocimiento automático, como contraseñas o tarjetas de identificación pueden ser robados, falsificados u olvidados. Contrariamente a esto, una característica biométrica debe ser universal, única, permanente, mensurable y de alto rendimiento, aceptable para los usuarios y debe ser complicado tanto como sea posible de eludir. Hoy nos encontramos con sistemas de reconocimiento de huella digital en teléfonos móviles, ordenadores portátiles y puertas. Con el aumento de la difusión de los sistemas biométricos, el rendimiento del reconocimiento en el campo de la ciencia forense también ha mejorado durante los últimos años.

Los sistemas biométricos han ayudado a esclarecer muchos casos, por poner algún ejemplo, a identificar a los dos sospechosos de los atentados de Boston en 2013 [119] y también para identificar a un sospechoso de una serie de robos en gasolineras en los Países Bajos [12]. En este último caso, el reconocimiento de orejas jugó un papel importante para la identificación del sospechoso.

La oreja como una característica biométrica durante mucho tiempo ha sido reconocida como un medio valioso para la identificación personal de los investigadores criminales. El criminólogo francés Alphonse Bertillon fue el primero en investigar el potencial para la identificación humana de las orejas hace más de un siglo, en 1890 [9]. En 1893 Conan Doyle publicó un artículo en el que describe las particularidades de las orejas de un grupo seleccionado y argumenta que la forma de la oreja, al igual que el rostro, refleja propiedades del carácter de una persona [10]. En sus estudios en 1906, Richard Imhofer sólo necesitó cuatro características diferentes para distinguir 500 orejas [198]. Más tarde, en 1964 el policía americano Alfred Iannarelli recogió más de 10.000 imágenes de orejas y determinó 12 características para identificar un sujeto [235]. Iannarelli también llevó a cabo estudios sobre gemelos y trillizos, donde descubrió que las orejas son aún únicas entre los sujetos genéticamente idénticos. Estudios científicos confirmaron posteriormente la suposición de Iannareli, que la forma externa de la oreja es única [146], que además es estable a lo largo de la vida humana [18].

En numerosos trabajos de investigación a partir de la última década, se pudo demostrar que el rendimiento del reconocimiento biométrico alcanzado con sistemas automatizados de reconocimiento de orejas es competitivo en comparación a los sistemas de reconocimiento facial [6]. Lei et al. también han demostrado que el oído externo se puede utilizar para la clasificación de género [154].

En muchos casos criminales no hay mayor evidencia que un vídeo en el que vemos al autor cometer un delito. Los investigadores criminales suelen tratar de combinar la información procedente de los testigos o de la víctima con lo obtenido de vídeos. La identificación biométrica es una herramienta útil para esta tarea, sin embargo, las condiciones no controladas en vídeos siguen siendo un entorno difícil para todos los sistemas de identificación automática. Las dificultades típicas de identificación automática en este caso de uso particular, son, por ejemplo, variaciones en la pose, escenas poco iluminadas, compresión de imágenes, falta de definición y ruido.

Los delincuentes, que son conscientes de la presencia de una cámara evitan frecuentemente mirar directamente a ella y algunas veces utilizan sombreros para cubrir sus rostros. Esto significa que los investigadores a menudo tienen que trabajar con imágenes de perfil que además puede ser parcial o totalmente ocluido. En estos escenarios, el reconocimiento de orejas puede ser un valioso aporte a los sistemas de reconocimiento facial existentes para identificar al sospechoso.

Mientras el vídeo contenga un solo frame, donde la oreja del sujeto sea claramente visible desde uno de los ángulos que coinciden con los puntos de vista de referencia, la identificación automática tiene una oportunidad de tener éxito. En la práctica, sin embargo, las imágenes limpias son raramente el caso. Además, la resolución de las cámaras de vigilancia en relación con la región supervisada puede ser mala y podemos encontrar degradaciones, como el desenfoque.

En un pequeño estudio, que fue realizado en la Universidad de Salamanca, entre septiembre 2013 y julio de 2014, tratamos de estimar la probabilidad media de que la oreja sea visible en público. También hicimos notas sobre el tipo de oclusión y el clima. Se ha observado que la oreja era totalmente visible en el 46 % de los casos de un total de 5.431 observaciones. Al mismo tiempo, la probabilidad de oclusión es altamente dependiente del género. Considerando que las orejas de las mujeres sólo eran plenamente visibles en 26,03 % de los casos, las orejas de los hombres eran totalmente visibles en 69,68 % de los casos. Más detalles sobre este estudio se pueden encontrar en el Apéndice A.

1.5.1. Identificación forense usando imágenes de orejas

En la identificación forense, los sistemas de identificación biométricos se utilizan para la recuperación de los candidatos más probables a partir de una gran base de datos. La fiabilidad de estos sistemas no sólo se reduce por las variaciones de

pose y oclusiones, sino también por el contraste y compresión dando como resultado imágenes de baja calidad. El uso de sistemas de identificación automáticos que utilizan imágenes 3D puede resultar en estimaciones más precisas, ofreciendo la posibilidad de ajustar la pose de referencia a la postura de la imagen de entrada. Tal estimación también podría incluir la forma de la oreja, a fin de incrementar la precisión, especialmente en los casos en que sólo el perfil de los sospechosos está disponible. En tal escenario, el reconocimiento de orejas es un recurso valioso como alternativa a los descriptores faciales lo que permitiría a los expertos forenses realizar su trabajo incluso con imágenes parciales.

Tan pronto como tenemos disponible una lista de los candidatos más probables, la evidencia puede ser recogida de forma manual con la superposición de trazados, donde se miden las distancias precisas entre puntos de referencia en un par de imágenes. Debe asegurarse de que sólo los puntos idénticos se comparan y, en caso de variación de pose, se necesitan dos imágenes con la misma pose. Posteriormente, los analistas verifican que tan bien los contornos de las dos imágenes coinciden. Al analizar imágenes de rostros con este método, el analista también puede investigar la simetría de ambos lados del rostro por medio de dos imágenes diferentes [81]. Las técnicas descritas anteriormente actualmente se aplican principalmente en imágenes faciales, pero pueden, en principio, ser utilizadas para cualquier tipo de imágenes incluyendo imágenes de orejas.

1.6. Objetivos

Este trabajo tiene por objetivo principal la exploración de nuevas técnicas para el reconocimiento de orejas, especialmente el uso de redes neuronales profundas. Nos centramos en imágenes 2D y no se limita al proceso de reconocimiento. Con estos antecedentes investigamos posibilidades de combinar diferentes perspectivas de las vistas de un objeto explotando el hecho de que la profundidad y la información de textura podrían ser útiles en el proceso de reconocimiento.

Analizamos aún más las propiedades estadísticas de los píxeles de las distintas imágenes y proponemos un método genérico para crear representaciones binarias para una técnica de búsqueda más eficiente. Aplicamos estos vectores de características binarios en un enfoque de búsqueda secuencial, vectores que se utilizan para la creación de una pequeña lista de los candidatos más probables.

Se presenta un framework para el reconocimiento de objetos basado en vistas en la que una imagen de consulta se representa como un gráfico, un gráfico que codifica la descomposición jerárquica de una silueta en partes primitivas. Tales valores primitivos surgen de un etiquetado de las singularidades a lo largo de la forma del objeto. La descomposición de la imagen se presenta en detalle en el capítulo 3, junto con un nuevo algoritmo para calcularla que mejora la precisión bajo una pequeña deformación en las formas debido a ruido o cambios en los puntos de vista. Dada una base de datos de distintas vistas de los objetos, el primer problema se formulará en cómo seleccionar un pequeño número de puntos de vista que representen los posibles candidatos para detectar e identificar dicho objeto.

El capítulo 2 resume el trabajo relacionado sobre algunos algoritmos de indexación y búsqueda de objetos e introduce algunas estrategias para realizar la selección de estos candidatos, como lo propuesto por Choras [161], identificando las limitaciones de este enfoque y explorando una serie de soluciones para superarlos. En particular, se desarrolló un eficiente algoritmo de acumulación de votos con el fin de encontrar una solución óptima al problema de las múltiples soluciones propuestas por distintas aproximaciones. En la etapa de verificación, cada uno de los candidatos se hace coincidir con la consulta dando una medida de similitud que se utiliza para clasificar a los posibles candidatos. El problema de encontrar la concordancia exacta entre la entrada y las posibles respuestas se convierte en un problema de la mejor asignación posible o mejor porcentaje de reconocimiento ya que se combinan distintos algoritmos tanto de detección e identificación en el que las correspondencias de objetos deben satisfacer un conjunto de restricciones.

Los objetivos del presente trabajo de investigación se pueden resumir en las siguientes preguntas planteadas:

- ¿Cómo puede ser detectada de forma automática la oreja en imágenes estáticas y vídeo?
- ¿Cómo pueden las imágenes ser normalizadas con respecto a la rotación y escala?
- ¿Es posible combinar los datos de diferentes perspectivas con el fin de obtener un mejor descriptor para alcanzar un mejor rendimiento?
- ¿Cómo pueden las orejas ser representadas en plantillas para habilitar una búsqueda más rápida en conjuntos de datos grandes?
- ¿Es posible encontrar automáticamente categorías de orejas?

Como una extensión de nuestros objetivos, se planteó el desarrolló de un módulo de reconocimiento de orejas como parte de un sistema de reconocimiento multimodal de rostros y orejas. Este sistema se evalúa y se prueba utilizando nuestro propio conjunto de datos; exploramos las virtudes y limitaciones del reconocimiento en este escenario y señalamos la orientación futura de este tipo de sistemas.

1.6.1. Estructura

El presente documento esta dividido en cuatro partes. Un resumen del proyecto que se desarrolló en el contexto de este trabajo, incluyendo la explicación de los requisitos del sistema, el sistema de captura de imágenes, el flujo de trabajo de los servicios de reconocimiento biométrico y algunas observaciones finales sobre el rendimiento general del mismo.

Comenzamos con una visión general sobre el estado del arte en el apartado 2; la estructura de esta segunda parte representa la revisión literaria sobre los avances en los procesos de detección, métodos de extracción y reconocimiento en el campo del reconocimiento de orejas. El tercer capítulo titulado descripción del sistema expone el detalle de los algoritmos utilizados a lo largo de la investigación, a fin de profundizar en el trabajo llevado a cabo para detectar y reconocer la oreja, donde, proponemos una técnica de ventana deslizante usando un marco de detección circular, algoritmo que evaluamos con respecto a su robustez frente a rotaciones, para posteriormente integrar otros algoritmos como redes neuronales convolucionales y la transformación de imagen por rayos. En la sección cuarta se presentan las conclusiones y trabajo futuro. El apéndice A expone un estudio empírico sobre la posibilidad de encontrar la oreja visible en el mundo real. El apéndice B sigue de forma general las directrices base que se proponen en la norma ISO/IEC SC 37 SD11 sobre los trabajos en el área biometrica [1]. Finalmente el C presenta el entorno de trabajo del sistema.

1.6.2. Preguntas de Investigación

Las siguientes preguntas se derivan de las cuestiones que se produjeron durante la investigación en curso. Dichas preguntas marcan la línea base, que ha seguido el trabajo de investigación.

1. **¿Cual es el estado del arte actual en el proceso de reconocimiento de orejas?** Para definir esta serie de preguntas, se ha elaborado un resumen sobre el estado del arte en el proceso de reconocimiento de personas a través de imágenes de las orejas. El resumen consiste en tres partes principales. En la primera parte, le damos una visión general a las bases de datos disponibles al público que se pueden utilizar para evaluar los sistemas de reconocimiento de orejas. A continuación se describen los diferentes enfoques para la detección en imágenes 2D y 3D y se compara su precisión. Posteriormen-

te estudiamos los sistemas de reconocimiento para dar una visión completa de los diferentes enfoques y su rendimiento. Esta compilación sirve como base para el trabajo realizado. Un resumen de los trabajos más recientes sobre el reconocimiento de orejas se proporciona como un último apartado en este capitulo.

2. **¿Cómo se puede detectar la oreja de forma automática en imágenes?** Una segmentación fiable es importante para cualquier sistema que intente realizar reconocimiento. En este trabajo, analizamos las características geométricas de la oreja y la forma en que se puede utilizar para la segmentación. Especial atención se fija en la fusión de textura y profundidad y la robustez sobre variaciones en la pose.

   Para la detección de orejas, las técnicas del estado del arte del reconocimiento de rostros, como las características Haar [189] dieron una exactitud satisfactoria en el proceso de detección, siempre y cuando la configuración de la imagen capturada sea controlada cuidadosamente y la calidad de imagen sea suficiente. Altos niveles de exactitud se pueden lograr utilizando el algoritmo LBP (implementación original según lo sugerido por Ahonen et al. [227]). También es posible detectar la oreja con el método de Regresión de Pose en Cascada [21] a partir de imágenes especialmente segmentadas, es decir, imágenes de perfil del rostro. Con el fin de asegurarse de que la detección de la oreja en la imagen 2D se realice lo suficientemente bien, la oreja no debe ser menor a 50*x*80 píxeles.

   Las orejas en imágenes de profundidad (3D) pueden ser segmentadas por la búsqueda de la estructura superficial única en la región exterior de la oreja. En [24] el enfoque de detección planteado por Zhou et al. [127] se extiende para detectar el oído externo en imágenes de perfil 3D bajo diferentes rotaciones en el plano. Debido a la proyección de la ROI *(Region of Interes)* de

coordenadas cartesianas a coordenadas polares, la exactitud de la detección disminuye. Pflug [22] introduce un método de detección, en el que se reconstruye el contorno de la oreja mediante la combinación de las regiones con alta curvatura superficie y las líneas que forman el contorno de la oreja. Finalmente, se incluye un proceso novel de detección de orejas en imágenes y vídeo utilizando redes neuronales convolucionales, el cual se profundiza en lo siguientes capítulos.

3. **¿Cómo pueden las imágenes recortadas de la oreja normalizarse con respecto a la rotación y escala?** Con el fin de aplicar el reconocimiento de orejas en entornos más desafiantes, la región de interés debe ser normalizada con respecto a la rotación y la escala.

   La regresión de pose en cascada (CPR) es un enfoque para la normalización de rostros, originalmente propuesto por Dollar et al [182]. CPR optimiza una función de pérdida que trata de minimizar la diferencia entre las características basadas en los niveles del gris locales dentro de la elipse de la oreja. En lugar de la localización de un número de puntos de referencia definidos visualmente, CPR utiliza características débiles para la estimación de la orientación de la oreja. Teniendo en cuenta que tenemos un número suficiente de imágenes de entrenamiento, CPR puede ser optimizado para ser robusto frente a oclusiones parciales y diferentes poses.

   Utilizando CPR se ajusta una elipse alrededor de la oreja, donde el eje mayor de la elipse representa la mayor distancia entre el lóbulo y la hélice superior [21]. A continuación, se compensa escala y rotación ajustando el centro, la longitud del eje mayor y la inclinación de la elipse de modo que el eje mayor sea vertical y tenga una longitud fija.

Se demuestra que el rendimiento del reconocimiento usando CPR antes de extraer el vector de características es significativamente mayor sin normalización. También muestra que el algoritmo CPR recorta la región de la oreja con precisión para diferentes regiones de interés que representan diferentes configuraciones de captura. Obviamente, el beneficio del uso de este algoritmo aumenta con una variación mayor en rotación y escala en el conjunto de datos.

4. **¿Qué impacto tiene la degradación de las imágenes sobre el desempeño de los sistemas de reconocimiento de orejas?** El rendimiento de cada sistema biométrico es dependiente de la calidad de las imágenes de entrada. La cuantificación de la calidad de la imagen, sin embargo, siempre depende de la situación. Hemos llevado a cabo una serie de experimentos para aprender más sobre el impacto del ruido y la falta de definición en el desempeño de los sistemas de reconocimiento de orejas.

   Hemos generado una serie de imágenes degradadas y calculado la proporción de ruido con el fin de cuantificar las imágenes degradadas y como estas afectan el sistema; este valor es definido vía el error cuadrado medio entre una imagen de calidad optima y una imagen degradada. A continuación, se midió el descenso del rendimiento en los algoritmos de detección y reconocimiento.

5. **¿Es posible encontrar automáticamente categorías de imágenes de orejas?** La mayor parte de los esfuerzos de investigación en el reconocimiento de orejas se concentran en lograr un rendimiento alto de reconocimiento en los conjuntos de datos. El siguiente paso hacia un sistema operacional es proporcionar técnicas para una búsqueda rápida y eficiente en grandes bases de datos.

Se analizan los espacios de características como la textura con respecto a las tendencias de agrupación que podrían ser explotadas para reducir el número de candidatos en una búsqueda 1 : N. Se crean entonces subespacios utilizando tanto métodos lineales como no lineales y se estudian estos subespacios para determinar las tendencias de agrupamiento. Se aplican diferentes métricas para estimar la bondad de las soluciones de clustering y mostrar, que los subespacios de características pueden ser organizados como grupos convexos utilizando *K-means*.

Se demuestra que la agrupación utilizando *K-means* para orejas 2D, PCA (*Principal Components Analysis*) para la proyección del subespacio y LPQ *(Local Phase Quantisation)* para la función de la textura es posible. Para esta configuración particular, el espacio de búsqueda se puede reducir a menos de 50 % de la base de datos con la probabilidad de 99, 01 % de que la identidad correcta este contenida en el conjunto de datos reducido. También se muestra que una búsqueda que se extiende a un máximo de tres grupos adyacentes obtiene un mejor rendimiento que una búsqueda en un solo clúster. Las clases dependen del tono de piel de los sujetos, y también de los ajustes al momento de la captura de las imágenes del conjunto de datos. También observamos que los vectores de características con un alto rendimiento en las clasificación o categorización de las imágenes no necesariamente alcanzan altas tasas de rendimiento en el reconocimiento.

De intentar responder a las preguntas planteadas se deriva el proponer un sistema en el que las distintas vistas de los objetos se descomponen en relaciones jerárquicas. Estos valores primitivos y sus relaciones se representan como vectores de características y se utilizan como entrada y entrenamiento para los algoritmos de coincidencia y reconocimiento. En las secciones siguientes se expone en profundidad la estrategia seguida para cumplir con el objetivo de la investigación.

2

Biometría de las orejas: estudio sobre detección, características y métodos de reconocimiento

El presente capítulo ofrece una visión elaborada del estado del arte del proceso de reconocimiento de orejas. Cuando se puso en marcha este proyecto se planteó la intención de responder a la pregunta de investigación ¿cuál es el actual estado del arte en el reconocimiento de orejas? cuestión que se intenta responder a lo largo del presente texto. Presentamos una visión general de las bases de datos disponibles y comparamos una gran selección de trabajos previos sobre segmentación, detección y reconocimiento con respecto a los enfoques aplicados y sus indicadores de desempeño. Concluimos con una sección que describe los retos futuros en el campo.

La posibilidad de identificar a las personas por la forma de su oído externo fue planteada por vez primera por el criminólogo francés Bertillon, y refinada por el oficial de policía Americano Iannarelli [235], quien propuso un primer sistema de reconocimiento de orejas basado en sólo siete características. La estructura detallada de la oreja no sólo es única, sino también permanente, su apariencia no cambia en el transcurso de la vida humana. Además, la adquisición de imágenes de orejas no requiere necesariamente la cooperación de una persona, por lo que es, considerada como una técnica no intrusiva.

Debido a estas cualidades, el interés por los sistemas de reconocimiento de orejas ha crecido de manera significativa en los últimos años. En este capítulo, categorizamos y resumimos enfoques para la detección y reconocimiento en imágenes 2D y 3D. A continuación, ofrecemos una visión de las probables futuras investigaciones en el campo, en el contexto de la vigilancia inteligente y análisis de imágenes forenses, que consideramos que es la aplicación más importante en un futuro cercano.

2.1. Introducción

Como existe una necesidad cada vez mayor para autenticar automáticamente a los individuos, la biometría ha venido siendo un campo de investigación activo en el transcurso de la última década. Los medios tradicionales de reconocimiento automático, como contraseñas o tarjetas de identificación, puede ser robadas, falsificadas, u olvidados. Las características biométricas, por otro lado, son universales, únicas, permanentes, y medibles.

El aspecto característico del oído externo humano (oreja) está formado por la hélix exterior, la antihélix, el lóbulo, el trago, la antitrago, y la concha (ver figura 2.1.1). Las numerosas crestas y valles en su superficie sirven como resonadores acústicos. Para frecuencias bajas el pabellón refleja la señal acústica hacia el canal

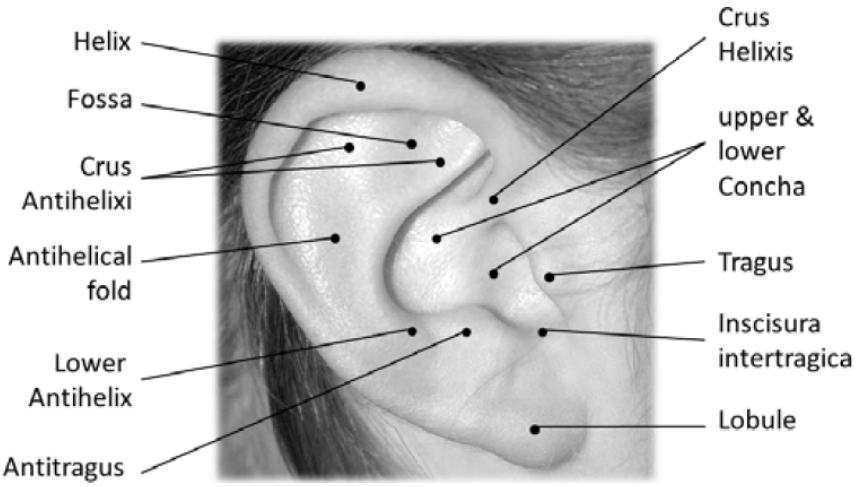


Figura 2.1.1: Características físicas de la oreja

auditivo. Las frecuencias altas reflejan las ondas sonoras provocando que las frecuencias circundantes sean omitidas. Además, el oído externo permite a los seres humanos percibir el origen de un sonido.

La forma de la oreja evoluciona durante el estado embrionario. Su estructura, por lo tanto, no es completamente al azar, pero todavía está sujeta a la segmentación celular. La influencia de factores aleatorios en la apariencia de la oreja se pueden observar mejor comparando la oreja izquierda y la oreja derecha de la misma persona. A pesar de que la izquierda y la oreja derecha muestran algunas similitudes, no son simétricas [3].

La oreja durante mucho tiempo ha sido reconocida como un medio valioso para la identificación personal por investigadores criminales. El criminólogo francés Alphonse Bertillon fue el primero en tomar conciencia del potencial uso para la identificación humana, hace más de un siglo [9]. En sus estudios sobre el reconocimiento a través del oído externo en 1906, Richard Imhofer necesitó sólo cuatro características diferentes para distinguir entre 500 orejas distintas [198]. A

partir de 1949, el oficial de policía estadounidense Alfred Iannarelli realizó el primer estudio a gran escala sobre el potencial discriminativo. Recogió más de 10.000 imágenes de orejas y determinó 12 características necesarias para identificar a una persona [235]. Iannarelli también llevó a cabo estudios sobre gemelos y trillizos, descubriendo que las orejas son aún únicas entre personas genéticamente idénticas. A pesar de que el trabajo de Iannarelli carece de una base teórica compleja, comúnmente se cree que la forma de la oreja externa es única. Los estudios realizados por Meijerman [145] y Shih [98] muestran que todas las orejas de las bases de datos investigadas poseen características individuales, que pueden ser utilizadas para distinguirse entre ellas. Debido a la falta de una base de datos suficientemente grande, estos estudios sólo se pueden ver como aproximaciones, no como evidencia, por la singularidad del oído externo.

La investigación sobre los cambios de apariencia ha mostrado que la oreja cambia ligeramente en tamaño cuando una persona envejece [57, 146]; esto es explicado por el hecho que con el envejecimiento la estructura microscópica de los cambios del cartílago de la oreja, reducen la elasticidad de la piel. Un primer estudio sobre el efecto de períodos cortos de tiempo en el reconocimiento de la oreja [164] muestra que la tasa de reconocimiento no se ve afectada por el envejecimiento. Debe ser, sin embargo, mencionado que la diferencia de tiempo en este experimento fue de sólo 10 meses, y por tanto, este hecho todavía está sujeto a profundizar las investigaciones sobre si el tiempo tiene un efecto crítico en la biométria de la oreja y por ende en los sistemas de reconocimiento o no.

La oreja puede ser capturada fácilmente desde la distancia, incluso si el sujeto no es del todo cooperativo. Esto hace al reconocimiento de orejas particularmente interesante para las tareas de vigilancia inteligente y para el análisis forense. Hoy en día la observación de características es una técnica estándar en la investigación forense y se ha utilizado como prueba en cientos de casos. La fuerza de esta evidencia ha sido, sin embargo, también puesta en duda por los tribunales por ejemplo en los Países Bajos [12]. Con el fin de estudiar la fortaleza de las impresiones de

la oreja como prueba, el proyecto forense de identificación (FearID) fue iniciado por nueve Institutos de Italia, el Reino Unido, y los Países Bajos en 2006. En su sistema de prueba, midieron un EER de 4 % y llegaron a la conclusión de que las impresiones de orejas pueden ser utilizadas como prueba en un sistema semiautomático [100]. La policía criminal alemana utiliza las propiedades físicas de esta característica biométrica en relación con otras propiedades basadas en apariencia para recolectar evidencia de la identidad de los sospechosos en las imágenes de las cámaras de vigilancia. La figura 2.1.1 ilustra los elementos más importantes y puntos de referencia de el oído externo, que son utilizados por la oficina federal de investigación criminal alemana para la identificación manual de sospechosos.

En este trabajo extendemos algunos estudios existentes sobre la biometría de las orejas, como por ejemplo lo expuesto por Pflug [20] o Ramesh et al [136] y otros varios autores que han centrado su trabajo en el procesamiento de imágenes 2D [134, 135, 159, 215]. Para el caso Abaza et al. [6] ha contribuido con un excelente estudio sobre el reconocimiento en marzo de 2010. Su trabajo abarca la historia de la biometría de las orejas, una selección de las bases de datos disponibles para la verificación de sistemas de reconocimiento a partir de imágenes 2D y 3D. Este trabajo modifica lo expuesto por Abaza et al. y Pflug et al. agregando lo siguiente:

- Un estudio de las bases de datos gratuitas y accesibles al público.
- Más de 30 publicaciones sobre la detección de y el reconocimiento de orejas publicadas en el periodo 2010-2015 y que no fueron discutidas en los estudios anteriores.
- Un punto de vista sobre los retos futuros para estos sistemas de reconocimiento con respecto a aplicaciones concretas.

2.2. Bases de datos disponibles

Con el fin de probar y comparar el rendimiento de detección o reconocimiento de un sistema de visión por computador, en general, y un sistema biométrico en particular, deben ser accesibles al público bases de datos de imágenes de suficiente tamaño. En este apartado, queremos presentar las bases de datos adecuadas para evaluar el rendimiento en cuanto a detección y reconocimiento se trata; bases de datos que pueden ser tanto descargadas libremente como licenciadas con un esfuerzo razonable.

2.2.1. Base de datos USTB

La Universidad de Ciencia y Tecnología en Beijing (*USTB - University of Science and Technology Beijing*) ofrece cuatro colecciones[1,2] de imágenes de orejas 2D y de imágenes de perfil para la comunidad de investigadores. Todas las bases de datos USTB están disponibles bajo licencia.

- **Base de datos I:** el conjunto de datos contiene 180 imágenes en total, las cuales fueron tomadas de 60 sujetos en tres sesiones entre julio y agosto de 2002. La base de datos sólo contiene imágenes de la oreja derecha de cada sujeto. Durante cada sesión, las imágenes fueron tomadas bajo diferentes condiciones de iluminación y con una rotación diferente. Los sujetos eran estudiantes y profesores de la USTB.

- **Base de datos II:** al igual que en la base de datos I, esta colección contiene imágenes de orejas derechas de los estudiantes y profesores de la USTB. Esta vez, el número de sujetos fue de 77 y fueron cuatro sesiones diferentes entre noviembre de 2003 y enero de 2004. Por lo tanto, la base de datos contiene 308 imágenes en total, que fueron tomadas en diferentes condiciones de iluminación.

[1] http://www1.ustb.edu.cn/resb/en/doc/Imagedb_123_intro_en.pdf
[2] http://www1.ustb.edu.cn/resb/en/doc/Imagedb_4_intro_en.pdf

- **Base de datos III:** en este conjunto de datos, 79 estudiantes y profesores de la USTB fueron fotografiados en diferentes poses entre noviembre y diciembre de 2004. Algunas de las orejas están ocluidas por el cabello. Cada sujeto giro su cabeza de 0 a 60 grados a la derecha y de 0 a 45 grados a la izquierda. Esto se repitió en dos días diferentes para cada materia, que se tradujo en 1.600 imágenes en total.

- **Base de datos IV:** consta de 25.500 imágenes de 500 sujetos tomadas entre junio de 2007 y diciembre de 2008, este es el mayor conjunto de datos de la USTB. El sistema de captación consistió en 17 cámaras y, fue capaz de tomar 17 imágenes del sujeto simultáneamente. Estas cámaras estaban distribuidas en un círculo alrededor del sujeto, quien fue colocado en el centro. El intervalo entre las cámaras era de 15 grados. Se pidió a cada voluntario que mirara hacia arriba, hacia abajo y hacia la altura de sus ojos, lo que significa que esta base de datos contiene imágenes en diferentes perspectivas. Hay que tomar en cuenta que esta base de datos sólo contiene una sesión por sujeto.

2.2.2. Bases de datos UND

La Universidad de Notre Dame (UND) ofrece una gran variedad de diferentes bases de datos de imágenes que se pueden utilizar para la evaluación del rendimiento biométrico. Entre ellas se encuentran cinco bases de datos que contienen imágenes 2D e imágenes 3D de profundidad que son adecuados para la evaluación de sistemas de reconocimiento de orejas. Todas las bases de datos de la UND están disponibles bajo licencia [3].

- **Colección E:** 464 imágenes del perfil derecho de 114 sujetos, capturadas en 2002. Para cada usuario, se tomaron entre 3 y 9 imágenes en diferentes días, distintas poses y condiciones de iluminación.

[3] http://cse.nd.edu/cvrl/CVRL/Data_Sets.html

- **Colección F:** 942 imágenes 3D y las correspondientes imágenes de perfil 2D, tomadas de 302 sujetos, capturadas en 2003 y 2004.
- **Colección G:** 738 imágenes 3D y las correspondientes imágenes de perfil 2D a partir de 235 sujetos, capturadas entre 2003 y 2005.
- **Colección J2:** 1800 imágenes 3D y las correspondientes imágenes de perfil 2D a partir de 415 sujetos, capturadas entre 2003 y 2005 [192].
- **Colección NDOff-2007:** 7398 en 3D y las correspondientes imágenes 2D de 396 rostros. La base de datos contiene diferentes poses codificadas en el nombres de archivo [53].

2.2.3. WPUT DB

La Universidad de West Pommeranian de Tecnología *(WPUT - The West Pomeranian University of Technology)* ha recopilado una base de datos de orejas con el objetivo de proporcionar datos más representativos que las colecciones comparables[4] [62]. La base de datos contiene 501 sujetos de todas las edades y 2.071 imágenes en total. Para cada sujeto, la base de datos contiene entre 4 y 8 imágenes, que fueron tomadas en diferentes días y en diferentes condiciones de iluminación.

Los sujetos también utilizan decoraciones que comúnmente se utilizan en el día a día, como pendientes, audífonos, y además, algunas orejas se ocluyen con el cabello. En la figura 2.2.1, se muestran algunas imágenes de ejemplo. La presencia de cada uno de estos factores se codifica en los nombres de archivo de las imágenes. La base de datos se puede descargar libremente desde la URL dada.

[4] http://ksm.wi.zut.edu.pl/wputedb/

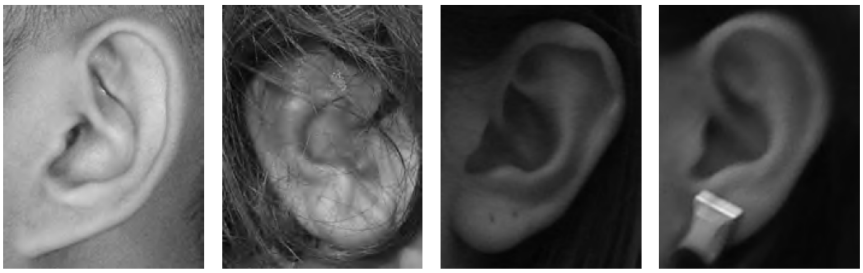


Figura 2.2.1: Ejemplo de la base de datos WPUT [62]. La base de datos contiene fotografias de orejas de diferente calidad y tomadas bajo diferentes condiciones de iluminación. Además incluye imágenes donde la oreja esta ocluida por el cabello y aretes

Figura 2.2.2: Ejemplo de las imágenes de la base de datos IIT Delhi [15]

2.2.4. IIT Delhi

La base de datos del Instituto Indio de Tecnologia IIT Delhi *(Indian Institute of Technology Delhi)* es proporcionada por la Universidad Politecnica de Hong Kong [5] [15]. Contiene imágenes de orejas que fueron recogidas entre octubre de 2006 y junio de 2007 en el Instituto de Tecnología Delhi, en Nueva Delhi, India (ver figura 2.2.2). La base de datos contiene 121 sujetos, y al menos tres imágenes fueron tomadas para cada uno en un ambiente interior, que significa que la base de datos consta de 421 imágenes en total.

[5] http://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Ear.htm

2.2.5. IIT Kanpur

La base de datos IITK *(Indian Institute of Technology Kanpur)* fue aportada por el Instituto Indio de Tecnología en Kanpur[6] [218]. Esta base de datos consta de dos subconjuntos.

- **Subconjunto I:** Este conjunto de datos contiene 801 imágenes del rostro de perfil recogidas de 190 sujetos. El número de imágenes adquiridas de cada individuo varía de 2 a 10.
- **Subconjunto II:** Las imágenes de este subgrupo fueron tomados de 89 individuos. Para cada sujeto se obtuvieron 9 imágenes con tres poses diferentes. Cada pose fue capturada en tres distintas escalas. Lo más probable es que todas las imágenes hayan sido tomadas el mismo día. No se indica si el subconjunto II contiene a las mismas personas que el anterior subconjunto.

2.2.6. ScFace

La base de datos SCface es proporcionada por la Universidad Técnica de Zagreb[7] [163], contiene 4.160 imágenes de 130 sujetos. El objetivo de la Universidad es proporcionar una base de datos, que sea adecuada para probar algoritmos bajo escenarios de vigilancia. Desafortunadamente, toda las imágenes de la cámara de vigilancia se tomaron en un ángulo frontal, de tal manera que las orejas no son visibles. Sin embargo, la base de datos también contiene un conjunto de fotografías de alta resolución, que muestra al sujeto en diferentes poses. Estas poses incluyen vistas del perfil derecho y perfil izquierdo, como se muestra en la figura 2.2.3. A pesar de que estas imágenes son probablemente inadecuadas para los estudios de reconocimiento de orejas, las fotografías de alta resolución podrían ser utilizadas para evaluar la resistencia de un algoritmo al cambio de perspectiva.

[6] http://www.cse.iitk.ac.in/users/biometrics/
[7] http://www.scface.org/

Figura 2.2.3: Ejemplo de las imágenes de la base de datos SCFace [163]

2.2.7. Sheffield base de datos de rostros

Esta base de datos era conocida anteriormente como la base de datos UMIST[8] y consta de 564 imágenes de 20 sujetos de raza y genero distinto. Cada sujeto se fotografía en una gama de diferentes poses, incluyendo la vista frontal y la de perfil.

2.2.8. YSU

La *Youngstown State University (YSU)* recoge un nuevo tipo de base de datos biométrica para la evaluación de sistemas de identificación forense [7]. Para cada uno de los 259 sujetos, se proporcionan 10 imágenes. Las imágenes se tomaron a partir de una secuencia de vídeo y muestran al sujeto en poses entre cero y 90 grados. Esto significa que la base de datos contiene imágenes de perfil recto y una vista frontal del rostro de cada individuo. También contiene bocetos dibujados a mano de 50 personas seleccionadas al azar desde un ángulo frontal. Sin embargo, esta parte de la base de datos no es de interés para un sistema de reconocimiento de orejas.

[8] http://www.sheffield.ac.uk/eee/research/iel/research/face

2.2.9. NCKU

La Universidad Nacional Cheng Kung de Taiwán ha recopilado una base de datos, que consta de 37 imágenes de 90 sujetos. Se puede descargar del sitio web[9] de la universidad. Cada sujeto fue fotografiado en diferentes ángulos entre -90 (perfil izquierdo) y 90 grados (perfil derecho) en pasos de 5 grados. En la figura 2.2.4 se muestran algunos algunos de ejemplo. Tal serie de imágenes se tomó en dos días para cada sujeto. Todas las imágenes fueron tomadas en las mismas condiciones de iluminación y con la misma distancia entre el sujeto y la cámara.

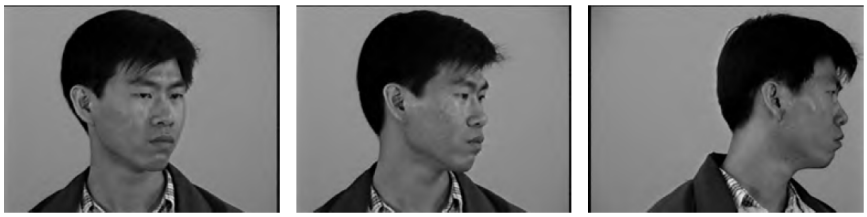


Figura 2.2.4: Ejemplo de las imágenes de la base de datos NCKU, mostrando al mismo sujeto desde diferentes ángulos

Esta información se recogió originalmente para el reconocimiento de rostros, algunas de las orejas están parte o totalmente ocluida por el cabello, lo que representa un reto para los enfoques de detección. En consecuencia, sólo un subconjunto de esta base de datos es adecuada para el reconocimiento de orejas.

2.2.10. Conjunto de datos UBEAR

El conjunto de datos presentado por Raposo et al. [203] contiene imágenes de la oreja izquierda y derecha de 126 sujetos. Las imágenes fueron tomadas bajo diferentes condiciones de luz y no se le pidió a los sujetos que se retiraran el cabello, joyas u otro tipo de decoración antes de tomarlas.

[9] http://robotics.csie.ncku.edu.tw/Databases/FaceDetect_PoseEstimate.htm

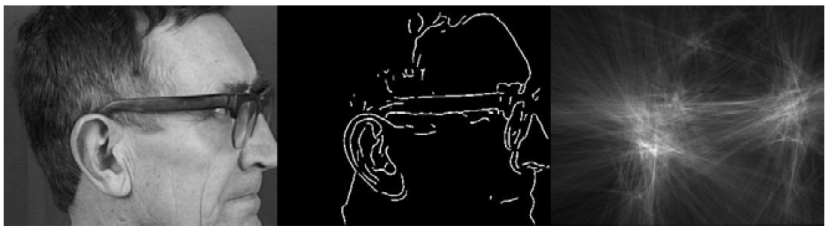


Figura 2.2.5: Imagen original, bordes y transformación de Hough

Figura 2.2.6: Imagen original y transformación por rayos

Las imágenes fueron recortadas de vídeo, lo que muestra al sujeto en diferentes poses, por ejemplo mirando hacia la cámara, hacia arriba o hacia abajo. A mayores se incluye en la base de datos la información de la posición exacta de la oreja, lo que hace a este conjunto particularmente conveniente para estudiar la exactitud de los algoritmos de detección e investigar y/o analizar el rendimiento del reconocimiento independientemente de cómo se detecte la oreja.

2.3. Detección de orejas

En esta sección se resume el estado del arte en la detección automática de orejas en imágenes 2D y 3D respectivamente. Básicamente todos los enfoques de detección se basan en las propiedades de la morfología física de las orejas, como la aparición de ciertos bordes característicos o patrones en su forma. Las tablas 2.3.1 y 2.3.2 ofrecen una breve visión general de los métodos de detección 2D y 3D.

Tabla 2.3.1: Métodos de detección 2D

Abaza et al. [5]	Cascaded Adaboost	940	2D	88,72 %
Ansari and Gupta [205]	Edge Detection and Curvature Estimation	700	2D	93,34 %
Alvarez et al [140]	Ovoid Model	*NA*	2D	*NA*
Arbab-Zavar & Nixon [37]	Hough Transform	942	2D	91 %
Arbab-Zavar & Nixon [38]	Log-Gabor Filters and Wavelet Transform	252	2D	88,4 %
Attarchi et al. [206]	Edge Detection and Line Tracing	308	2D	98,05 %
Chen & Bhanu [87]	Template Matching with Shape Index Histograms	60	2D	91,5 %
Cummings et al. [11]	Ray Transform	252	2D	98,4 %
Islam et al. [213]	Adaboost	942	2D	99,89 %
Jeges & Mate [72]	Edge Orientation Pattern	330	2D	100 %
Kumar et al. [17]	Edge Clustering and Active Contours	700	2D	94,29 %
Liu & Liu [90]	Adaboost and Skin Color Filtering	50	2D	96 %
Prakash & Gupta [220]	Skin Color and Graph Matching	1780	2D	96,63 %
Shih et al. [98]	Arc-Masking and AdaBoost	376	2D	100 %
Yan & Bowyer [192]	Concha Detection and ActiveContours	415	2D	97,6 %
Yuan & Mu [151]	CAMSHIFT and a Contour Fitting	*Video*	2D	*NA*

Tabla 2.3.2: Métodos de detección 3D

Publicación	Método de detección	Tamaño database	Tipo	Precisión
Chen & Bhanu [88]	Shape model and ICP	700	3D	87,71 %
Chen & Bhanu [86]	Helix Shape Model	213	3D	92,6 %
Zhou et al. [127]	Histogram of Categorized Shapes	942	3D	100 %
Prakash & Gupta [219]	Connectivity Graph	1604	3D	99,38 %

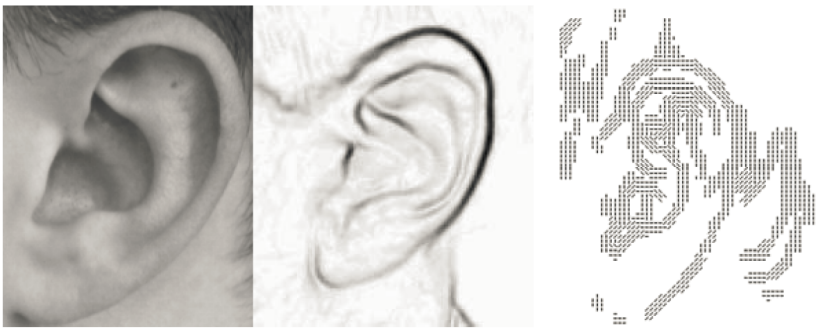


Figura 2.3.1: Imagen original, bordes mejorados y su correspondiente modelo orientado por los bordes

La tabla 2.3.2 contiene los algoritmos de localización 3D, mientras que la tabla 2.3.1 los algoritmos diseñados para la detección de orejas en imágenes 2D. Las figuras 2.2.5, 2.2.6 y 2.3.1 representan ejemplos de distintas técnicas de detección de orejas. Chen y Bhanu proponen tres enfoques diferentes para la detección de la oreja. En el enfoque expuesto por Chen et al. [87] entrenan un clasificador, que reconoce una distribución específica de índices, que son característicos de la superficie de la oreja. Sin embargo, este enfoque sólo funciona en imágenes de perfil y es sensible a cualquier tipo de rotación, escala y variación en las poses.

Más adelante en sus aproximaciones detectan áreas con gran curvatura local con una técnica que llamaron paso de magnitud de bordes [86]. A continuación, generan una plantilla que contiene la forma típica del helix exterior y el anti-helix de la oreja colocada de tal forma que agrupa las líneas de su contorno. Chen et al [88] estrechan el número de posibles candidatos detectando primero la región de la piel antes de realizar la comparación de plantillas del helix. Fusionando el color y la información de la curvatura, la tasa de detección podría verse incrementada hasta 99, 3 % en la base de datos UCR y hasta 87, 71 % sobre la colección F y G de la base de datos UND. El conjunto de la UCR no es accesible al público y por tanto no se cubre en la sección 2.2. Para una descripción mayor de su contenido se puede consultar el trabajo de Abaza et al. [6].

Otro ejemplo de detección de orejas utilizando el contorno es descrito por Attarchi et al. [206]. Ellos localizan el contorno exterior mediante la búsqueda del borde más largo conectado en la imagen pre-procesada. Seleccionando los puntos superiores, inferiores, e izquierdos de los limites detectados, forman un triángulo para posteriormente utilizar el centro como punto de referencia para la alineación de la imagen. Ansari et al. también utilizan un detector de bordes en la primera etapa de su enfoque de localización [205]. Los bordes son separados en dos categorías, llamadas, convexa y cóncava. Se eligen los bordes convexos como candidatos para representar el contorno exterior. Finalmente, el algoritmo conecta los segmentos curvos y selecciona la figura que encierra el área más grande lo que a efectos del proceso representará el contorno exterior de la oreja. Cabe señalar que la base de datos IITK y USTB II ya contienen imágenes cortadas de orejas. Por lo tanto, se puede cuestionar, si las tasas de detección de 93, 34 % y 98, 05 % pueden ser reproducidas en condiciones realistas.

Un enfoque reciente sobre detección de orejas en 2D que utiliza bordes es descrito por Prakash y Gupta [219]. Combinan la segmentación de la piel y la categorización de los bordes en convexos y cóncavos. Después, los bordes en la región

de la piel seleccionada se descomponen en segmentos. Estos segmentos se componen para formar un gráfico de conectividad de bordes. A partir de este gráfico se calcula el casco convexo de todos los bordes, que se cree que pertenecen a la oreja. La región cerrada es entonces etiquetada como la región de interés. En contraste con Attarchi et al. [206], Prakash y Gupta demuestran la viabilidad de la detección basada en bordes *(edge-based)*, probando su teoría en imágenes de perfil completo, donde lograron una tasa detección de 96, 63 % en un subconjunto de la colección UND-J2. Prakash et al. [219] proponen la misma conectividad basada en bordes para el reconocimiento de orejas en imágenes 3D. En lugar de los bordes, utilizan discontinuidades en el mapa de profundidad para extraer la imagen de bordes inicial y luego extraer el gráfico de conectividad. En sus experimentos, utilizan las representaciones 3D del mismo subconjunto como en [220] y reportan una tasa de detección de 99, 38 %. Además muestran que la tasa de detección de su enfoque basado en este gráfico no es influenciado por la rotación y la escala.

Jedges y Mate proponen otro enfoque de detección basado en bordes *(edge-based)*, que probablemente esté inspirado en técnicas de reconocimiento de huella dactilar. Entrenan un clasificador con patrones de orientación que previamente han sido calculados a partir de imágenes de orejas. Al igual que otros clasificadores, su método no es robusto frente a rotación y escala. Además, el clasificador es probable que falle bajo grandes variaciones en la pose, ya que esto afecta a la apariencia del patrón de orientación.

Islam et al. [214] y Abaza et al. [5] utilizan clasificadores débiles basados en ondas Haar con conexión AdaBoost para localizar la oreja. Según Islam et al., el entrenamiento de los clasificadores tarda varios días, sin embargo una vez que el clasificador está configurado, la detección de la oreja es rápida y eficaz. Abaza et al. utilizan una versión modificada de AdaBoost y reportan una fase de entrenamiento más corta. La eficacia de su enfoque se demostró en evaluaciones a cinco diferentes bases de datos. También se incluyen algunos ejemplos exitosos de detecciones en imágenes obtenidas de Internet. Mientras la pose del sujeto no cambie, los cla-

sificadores débiles son adecuados para imágenes que contienen más de un sujeto. Dependiendo del conjunto de prueba seleccionado Abaza et al. alcanzan una tasa de detección de entre 84 % y 98, 7 % en la base de datos de rostros Sheffield. En promedio, su enfoque detectó con éxito el 95 % de todas las orejas.

Yan y Bowyer desarrollaron un método de detección que fusiona imágenes de rango y sus correspondientes imágenes a color en 2D [192]. Su algoritmo comienza localizando la concha y a continuación, utiliza contornos activos para determinar el límite exterior de la oreja. La concha sirve como el punto de referencia para definir la forma inicial del modelo de contorno activo. Incluso aunque la concha es fácil de localizar en imágenes de perfil, puede estar ocluida si la pose de la cabeza cambia o si el sujeto lleva un arete o audífonos. En sus experimentos Yan y Bowyer sólo utilizan imágenes de orejas con oclusiones mínimas donde la concha es visible; por lo tanto podría no ser probado ni refutado si su enfoque es capaz de manera fiable de detectar las orejas si la concha esta ocluida.

Yuan y Mu desarrollaron un método para el seguimiento de la oreja en tiempo real en secuencias de vídeo aplicando continuamente el algoritmo adaptativo *Mean Shift (CAMSHIFT)* [151]. El algoritmo CAMSHIFT se utiliza con frecuencia en aplicaciones de seguimiento del rostro y se basa en concordancia de regiones y modelos del color de la piel. Para lograr precisión en la segmentación de la oreja, se aplica, el método de ajuste de contorno basado en modelos de forma activos modificados, que han sido propuestas por Alvarez et al. [140]. Yuan y Mu reportan una tasa de detección del 100 %, sin embargo la base de datos de prueba solamente consistía de dos sujetos. Sin embargo, su enfoque parece ser muy prometedor para aplicaciones de vigilancia pero necesita ser evaluado en escenarios más realistas. Shih et al. determinan las orejas candidatas mediante la localización de los bordes en forma de arco en una imagen de bordes. Posteriormente los candidatos se verifican mediante el uso de un clasificador Adaboost. Su aproximación alcanza una tasa de detección del 100 % en un conjunto de datos, que consta de 376 imágenes de 94 sujetos.

Zhou et al. entrenan un modelo 3D con el fin de reconocer el histograma de los índices de la forma típica de la oreja [127]. De manera similar a los planteamientos de Abaza et al. e Islam et al., una ventana de deslizamiento de diferentes tamaños se mueve sobre la imagen. El descriptor de la oreja propuesto por Zhou et al. está construido a partir de histogramas índice concatenados, que se extraen de subbloques dentro de la ventana de detección. Para la detección real, un clasificador SVM *(Support Vector Machine)* es entrenado para decidir si una región de la imagen es la región de la oreja o no. Por lo que sabemos, éste es el primer enfoque de detección, que no requiere tener imágenes de textura además de imágenes de rango. Zhou et al. evaluaron su enfoque en imágenes de las colecciones UND obteniendo una tasa de detección del 100 %. Cabe señalar que este enfoque no fue evaluado bajo rotación, variaciones en la pose ni grandes oclusiones, pero bajo la impresión del buen rendimiento, creemos que este algoritmo es bastante interesante para futuras investigaciones.

Los métodos de detección basados en transformaciones de imagen tienen la ventaja de ser robustos frente a las rotaciones fuera de plano. Están diseñados para resaltar las propiedades específicas de la forma exterior de la oreja, que se observan en cada imagen donde la oreja es visible no importa la pose en que dicha oreja haya sido fotografiada. Arbab et al. [37] utilizaron la transformada de Hough para mejorar las regiones con una alta densidad de bordes. En las imágenes de perfil, una alta densidad de bordes se observa sobre todo en la región de la oreja (ver figura 2.2.2). Los autores reportan que la transformada de Hough reduce su eficiencia para la detección cuando las personas llevan gafas ya que el marco introduce bordes adicionales a la imagen. Esto ocurre especialmente en la región de los ojos y la nariz. El enfoque basado en la transformada de Hough para detección de orejas se evaluó en imágenes de la base de datos XM2VTS; el trabajo de Abaza et al. [6] expone una descripción detallada de la base de datos, donde se logró una tasa de detección del 91 %.

El enfoque de transformación de rayos propuesto por Cummings et al. [11] se ha diseñado para detectar la oreja en diferentes poses. Este algoritmo utiliza una especie de rayo de luz para escanear la imagen buscando las estructuras tubulares y curvas como la hélice exterior. El rayo simulado se refleja en las regiones brillantes y tubulares destacando estas regiones en la imagen transformada. Sin embargo, el rayo también destaca bordes rectos y los bordes de otros objetos, como el pelo y las gafas (ver figura 2.2.1). Usando este método Alastair et al. han logrado una impresionante tasa de reconocimiento de $98,4\,\%$ en la base de datos XM2VTS. Por lo tanto, el enfoque basado en la transformación de rayos de Cummings et al. supera a la transformada de Hough, muy probablemente porque es más robusto frente a factores de ruido tales como gafas o cabello.

Un enfoque reciente para la detección de orejas en imágenes 2D es descrito por Kumar et al. [17] quienes proponen extraer orejas de imágenes 2D utilizando imágenes de bordes y contornos activos. Evalúan su enfoque en una base de datos que contiene 100 sujetos con siete imágenes por sujeto. Un dispositivo especial se utilizó para recolectar los datos. Este dispositivo asegura que la distancia a la cámara es constante y que las condiciones de iluminación son las mismas para todas las imágenes. Esta aproximación reporta una tasa de detección de $94,29\,\%$.

Al poner en práctica estos algoritmos, podemos observar que la robustez frente al ruido como diferentes poses y oclusión es de vital importancia. Sin embargo, la mayoría de los métodos de detección descritos anteriormente no fueron probados en escenarios de oclusión realistas, como la oclusión del cabello, aretes, audífonos, etc. Una posible razón para esto puede ser la falta de bases de datos que contengan imágenes apropiadas, sin embargo este vacío ha sido llenado recientemente por diferentes grupos de trabajo, quienes han contribuido proporcionando conjuntos de datos apropiados (ver sección 2.2). Por otra parte, no hay investigaciones sobre el efecto de oclusión de la oreja en imágenes 3D.

2.4. Reconocimiento de orejas 2D

Cada sistema de reconocimiento de orejas consiste en una extracción de características y un paso de comparación del vector de características. En esta revisión de la literatura se divide el reconocimiento de orejas en cuatro subclases diferentes denominadas enfoques holísticos, locales, híbridos y enfoques estadísticos. En las tablas 2.4.1 y 2.4.2 todos los enfoques de reconocimiento 2D mencionados en este documento son resumidos en orden cronológico.

2.4.1. Descriptores holísticos

Otro enfoque, que ha ganado cierta popularidad es la transformación de campo de fuerza propuesta por Hurley [117]. El enfoque de transformación de campo de fuerza asume que los píxeles tienen una atracción mutua proporcional a sus intensidades e inversamente al cuadrado de la distancia entre ellos, algo como la ley universal de Newton de la gravedad. El campo de energía asociado toma la forma de una superficie lisa con un número de picos unidos por crestas (ver la figura 2.4.2).

Usando este método, Hurley et al. [117] lograron un rendimiento de rango-1 de más de 99 % en la base de datos de XM2VTS (252 imágenes). Basándose en estos resultados, Abdel-Mottaleb y Zhou utilizaron una representación en 3D del campo de fuerza para los puntos de extracción situadas en el pico de los campos de fuerza 3D [156]. Debido a que el campo de fuerza converge en el contorno de la oreja, los picos en la representación 3D básicamente representan la forma exterior de dicha oreja. El método de campo de fuerza es más robusto frente al ruido que otros detectores de bordes, como Sobel o Canny. Usando este enfoque, Abdel-Mottaleb y Zhou lograron un rendimiento de rango 1 de 87, 93 % en un conjunto de datos que consta de 103 imágenes de 29 sujetos.

Dong y Mu [114] añaden invariabilidad de pose a los bordes, que se extraen mediante el uso del método de campo de fuerza. Esto se consigue con el análisis discriminante de Fishier con kernel de espacio nulo *(NKFDA)*, que tiene la propiedad de representar relaciones no lineales entre dos conjuntos de datos. Dong y Mu realizaron experimentos en el conjunto de datos USTB IV. Antes de la función extracción, la región de la oreja fue recortada manualmente de las imágenes y se normalizó la pose. Para variaciones de pose de 30 grados reportan una tasa de reacondicionamiento del $72,2\,\%$. Para variaciones de 45 grados el rendimiento se redujo a $48,1\,\%$.

Una reciente publicación de Kumar y Wu [15] presenta un enfoque de reconocimiento de Orejas, que utiliza la información de los filtros Log-Gabor para codificar la estructura local de la oreja. La información codificada se almacena en forma de imágenes en escala de grises normalizada. En los experimentos, el enfoque Log-Gabor superó los enfoques basados en características aplicando campo de fuerza y la extracción de características basada en punto de referencia. Por otra parte, diferentes implementaciones de filtros Log-Gabor se compararon uno con otro dando como resultado que su rendimiento oscila entre $92,06\,\%$ y $95,93\,\%$, esto en una base de datos que contiene 753 imágenes de 221 sujetos.

La rica estructura del contorno de la oreja resulta en información de textura bastante específica, lo que puede medirse utilizando filtros Gabor. Wang y Yuan [243] extraen características de frecuencia locales utilizando una batería de filtros Gabor para luego seleccionar las características más distintivas utilizando un análisis discriminante general. En sus experimentos sobre la base de datos USTB II, compararon el impacto en el rendimiento de los diferentes ajustes de los filtros Gabor. Las diferentes combinaciones de orientación y escala en el conjunto de filtros fueron comparados entre sí y se encontró que ni el número de escalas ni el número de orientaciones tiene un impacto importante en el rendimiento.

Tabla 2.4.1: Métodos de reconocimiento de orejas en imágenes 2D parte 1

Publicación	Resumen	Cantidad sujetos	Imágenes	Precisión
Burge & Burger [158]	Vornoi Distance Graphs	NA	NA	NA
Yuan & Mu [149]	Full Space LDA with Outer Helix Feature Points	79	1501	86,76 %
Hurley [117]	Force Field Transform	63	252	99 %
Moreno et al. [41]	Geometric features with Compression Network	28	268	93 %
Yuizono et al. [232]	Genetic Local Search	110	660	99 %
Victor et al. [43]	PCA	294	808	40 %
Chang et al [131]	PCA	114	464	72,7 %
Abdel-Mottaleb & Zhou [156]	Modified Force Field Transfrom	29	58	87,9 %
Mu et al. [251]	Geometrical measures on edge images	77	308	85 %
Abate et al. [2]	General Fourier Descriptor	70	210	88 %
Lu et al. [155]	Active Shape Model and PCA	56	560	93,3 %
Yuan et al. [152]	Non-Negative Matrix Factorization	77	308	91 %
Arbab-Zavar et al. [39]	SIFT points from ear model	63	252	91,5 %
Jeges & Mate [72]	Distorted Ear Model with Feature Points	28	4060	5,6 %*EER*
Liu et al. [92]	Edge-based features from different views	60	600	97,6 %
Nanni & Lumini [147]	Gabor Filters and SFFS	114	464	80 %
Rahman et al. [168]	Geometric Features	100	350	87 %
Sana et al. [31]	Haar Wavelets and Hamming Distance	600	1800	98,4 %
Arbab-Zavar & Nixon [38]	Log-Gabor Filters	63	252	85,7 %
Choras [161]	Geometry of ear outline	188	376	86,2 %

Tabla 2.4.2: Métodos de reconocimiento de orejas en imágenes 2D parte 2

Publicación	Resumen	Cantidad Sujetos	Imágenes	Precisión
Dong & Mu [114]	Force Field Transform and NKFDA	29	711	75,3 %
Guo & Xu [247]	Local Binary Pattern and CNN	77	308	93,3 %
Naseem et al. [107]	Sparse representation	32	192	96,88 %
Wang et al. [249]	Haar Wavelets and Local Binary Patterns	79	395	92,41 %
Xie & Mu [252]	Locally Linear Embedding	79	1501	80 %
Yaqubi et al. [169]	HMAX and SVM	60	180	96,5 %
Zhang & Mu [97]	Geometrical Features, ICA and PCA with SVM	77	308	92,21 %
Badrinath & Gupta [80]	SIFT landmarks from ear model	106	1060	95,32 %
Kisku et al. [64]	SIFT from different Color Segments	400	800	96,93 %
Wang & Yuan [242]	Low-Order Moment Invariants	77	308	100 %
Alaraj et al. [157]	PCA with MLFFNNs	17	85	96 %
Bustard et al. [111]	SIFT Point Matches	63	252	96 %
De Marisco et al. [162]	Partitioned Iterated Function System (PIFS)	114	228	61 %
Gutierrez et al. [142]	MNN with Sugeno Measures and SCG	77	308	97 %
Wang et al. [245]	Moment Invariants and BP Neural Network	*NA*	60	91,8 %
Wang & Yuan [243]	Gabor Wavelets and GDA	77	308	99,1 %
Prakash & Gupta [218]	SURF and NN classifier	300	2066	2,25 %*EER*
Kumar et al. [17]	SIFT	100	700	95 %
Wang & Yan [253]	Local Binary Pattern and Wavelet Transform	77	308	100 %
Kumar & Wu [15]	Phase encoding with Log Gabor Filters	221	753	95,93 %

El rendimiento total del enfoque de Wang y Yuan fue de 99, 1 %. En un enfoque similar Arbab-Zavar y Nixon [38] midieron el rendimiento de los filtros Gabor en la base de datos XM2VTS donde reportaron un rendimiento de 91, 5 %. Una mirada más cercana a la respuesta del filtro Gabor mostró que los vectores de características estaban dañados por la oclusión u otros factores de ruido. Con el fin de superar esto, se propuso un método de comparación más robusto, que resultó en una tasa de reconocimiento mejorada del 97, 4 %.

Abate et al. [75] utilizan un descriptor de Fourier genérico para la representación de las características sin variación de rotación y escala. La imagen se transforma en un sistema de coordenadas polares para luego transformarse en un espacio de frecuencias. Con el fin de asegurarse que el centro de gravedad del sistema de coordenadas polares este siempre en la misma posición, las imágenes de las orejas tienen que estar alineadas antes de ser transformadas. La concha sirve de punto de referencia para la etapa de alineación, de tal manera que el centro siempre se encuentra en la región de la concha. El enfoque fue probado en un conjunto de datos propio de los investigadores, el cual contiene 282 imágenes en total. Las imágenes fueron tomadas en dos días diferentes y en diferentes poses. El rendimiento del descriptor de Fourier varía dependiendo del ángulo de la oreja desde el que se observe. Para variaciones de cero grados el rendimiento es del 96 %, pero si incluimos diferentes poses este se reduce a 44 % y 19 % para 15 y 30 grados respectivamente.

En el trabajo de Fooprateepsiri y Kurutach [196] se explotan los conceptos de transformación de traza a multi-resolución y la transformada de Fourier. Las imágenes de entrada de la base de datos CMU PIE son serializadas utilizando la transformación de traza y almacenados en un vector de características. La ventaja de la transformación de traza es que el vector de características resultante es invariante a rotaciones y escala. Además Fooprateepsiri y Kurutach muestran que su descriptor también es robusto contra variaciones en las poses. En total reportan un rendimiento de 97 %.

Sana et al. utilizan un selecto grupo de coeficientes Wavelet extraídos durante la compresión de la imagen con el algoritmo Haar-Wavelet para la representación de características [31]. Mientras aplica el nivel cuatro de la transformada Wavelet varias veces en la imagen de la oreja, para cada iteración se almacena uno de los coeficientes derivados en un vector de características. La precisión informada de su algoritmo es de 96 % y se alcanzó aplicando el algoritmo en el base de la base de datos IITK y en la base de datos Saugor que contiene 350 sujetos.

Un sistema de extracción de características llamada PIFS es propuesto por De Marisco et al. [162]. PIFS mide la auto-similitud en una imagen mediante el cálculo de traducciones afines entre sub-regiones similares de una imagen. Con el fin de hacer a su sistema robusto a la oclusión, De Marisco et al. dividieron la imagen de la oreja en secciones de igual tamaño. Si una sección esta ocluida, las otras deberían contener un conjunto de características suficientemente distintivo. De Marisco et al. pudieron demostrar que su enfoque es superior a otros métodos de extracción de características en virtud de la presencia de oclusión. Los experimentos se han realizado con el fin de evaluar el rendimiento del sistema en diferentes escenarios de oclusión. La base para estas pruebas fueron la UND colección E y los primeros 100 sujetos de la base de datos FERET. Si la oclusión se produce en la imagen de referencia, la precisión de reconocimiento se sitúa en 61 % en comparación al 40 % promedio de otros métodos de extracción de características, sin oclusión, el rendimiento es de 93 %.

Los momentos invariantes son una medida estadística para describir las distintas propiedades específicas de una forma. Wang et al. [245] componen seis vectores de características diferentes mediante el uso de siete momentos invariantes. También muestran que cada uno de los momentos invariantes es robusto frente a cambios de escala y rotación. Los vectores de características se utilizan como entrada para una red neuronal de retropropagación que es entrenada para clasificar los conjuntos de características. Basado en una base de datos propietaria de 60

imágenes de orejas, que reportan un rendimiento de 91,8 %. Wang y Yuan [242] comparan el carácter distintivo de los diferentes métodos de extracción de características en la base de datos USTB-I. Ellos comparan el rendimiento de descriptores de Fourier, transformación de Gabor, momentos invariantes y funciones estadísticas y llegan a la conclusión de que la más alta tasa de reconocimiento se puede lograr mediante el uso de momentos invariantes y la transformación de Gabor. Para ambos métodos de extracción de características Wang y Yuan obtienen un rendimiento del 100 %.

2.4.2. Descriptores locales

Scale-Invariant Feature Transform (SIFT) es un algoritmo conocido por ser una forma sólida de extracción de puntos clave incluso en imágenes con pequeñas variaciones en las poses y diferentes condiciones de luminosidad [65]. Los puntos de referencia *(landmarks)* SIFT contienen una medida para la orientación local; también pueden ser utilizados para estimar la rotación y traslación entre dos imágenes de orejas normalizadas. Bustard et al. mostraron que el algoritmo SIFT puede manejar variaciones en las poses hasta de 20 grados [111]. Sin embargo, no es algo trivial asignar un punto de referencia SIFT con su contraparte exacta, especialmente en presencia de variaciones.

En áreas de imagen altamente estructuradas, la densidad y la redundancia de los puntos clave SIFT son tan altas, que la asignación exacta no es posible. Por lo tanto los puntos de referencia tienen que ser filtradas antes de comenzar a realizar la comparación. Arbab-Zavar et al. [37], así como Badrinath y Gupta [80] entrenaron un modelo de puntos de referencia que sólo contiene un pequeño número de puntos no redundantes. Este modelo se utiliza para filtrar dichos puntos SIFT, los cuales fueron detectados inicialmente en la oreja de referencia. Teniendo los puntos clave filtrados es posible asignar cada uno de ellos con su contraparte en la imagen "modificada". La figura 2.4.1 muestra un ejemplo del algoritmo SIFT y

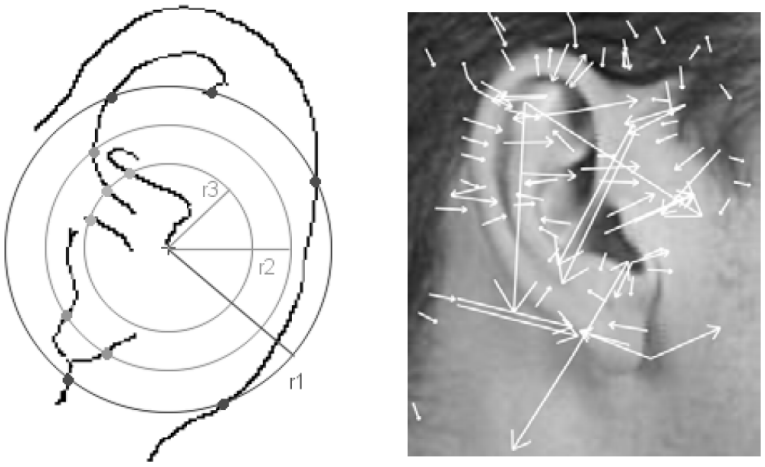


Figura 2.4.1: Círculos concéntricos [161], características SIFT [39]

de los puntos extraídos a partir de imágenes de orejas, características que fueron utilizadas como datos de entrenamiento para el modelo planteado en el trabajo de Arbab-Zavar et al. sus resultados pueden ser directamente comparados con la precisión demostrada por Bustard y Nixon [111], ya que ambos estudios utilizaron la base de datos XM2VTS para evaluación. Arbab-Zavar et al. registraron una tasa de desempeño del 91,5 %, mientras el enfoque más reciente de Bustard y Nixon mostró una tasa de precisión del 96 %. Utilizando la base de datos IIT Delhi Kumar et al. reportaron un tasa de aceptación genuina (GAR) de 95 % y tasa de aceptación falsa (FAR) del 0,1 % cuando se utilizan las características de puntos SIFT para reconocimiento.

Kisku et al. dirigen el problema de corregir los puntos de referencia asignados descomponiendo la oreja en diferentes segmentos de color [64]. Los puntos de referencia SIFT se extraen de cada segmento separadamente, lo cual reduce la posibilidad de asignar valores que no representen las mismas características. Usando este enfoque, Kisku et al. han registrado un desempeño de aproximadamente 97 %.

Un enfoque reciente de Prakash y Gupta [218] fusiona los puntos característicos obtenidos del algoritmo *Speeded Up Robust Features (SURF)* [85] de diferentes imágenes del mismo sujeto. Proponen utilizar varias imágenes de entrada del mismo sujeto para enrolar y almacenar todos los puntos característicos SURF en un vector de características fusionado. Este conjunto de características se utiliza para el entrenamiento de un clasificador del vecino más cercano *(nearest neighbor classifier)* para asignar dos puntos característicos correlacionados. Si la distancia entre dos puntos es menor que el umbral entrenado se considera que están correlacionados. La evaluación de este enfoque se llevó a cabo en la base de datos UND colección E y los dos subconjuntos de la base de datos IIT Kanpur. Prakash y Gupta pusieron a prueba la influencia de diferentes parámetros para obtener las características SURF y para el clasificador de vecino más cercano. Dependiendo del valor establecido en dichos parámetros el EER *(Equal Error Rate)* varía entre 6, 72 % y 2, 25 %. La tasa EER es la mejor descripción en un único valor de la tasa de error de un algoritmo.

2.4.3. Enfoques híbridos

El enfoque de Jedges y Mate es doble [72]. En una primera etapa generan un modelo promedio de bordes de un conjunto de imágenes de entrenamiento. Estos bordes representan el contorno exterior del helix, así como los contornos del antihelix, la fosa triangular y la concha. Posteriormente cada imagen es registrada deformando la oreja modelo hasta que se ajuste a los bordes reales que aparecen en la imagen de la oreja. Los parámetros de deformación, que fueron necesarios para la transformación, representan la primera parte del vector de características (*feature vector*). Este vector se completa con la adición de los puntos de características de las intersecciones entre un conjunto de ejes predefinidos y los principales bordes transformados. Los ejes describen el contorno único de la oreja.

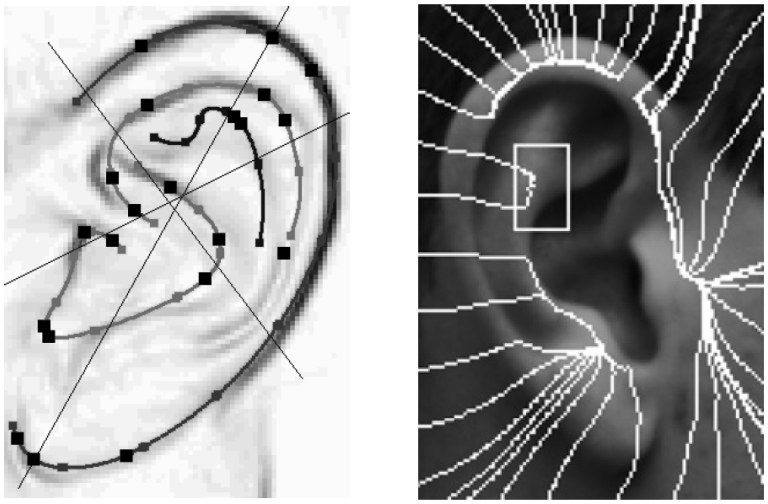


Figura 2.4.2: Contornos activos [72], campos de fuerza [117]

La figura 2.4.2 muestra los bordes mejorados en las imágenes con contornos que encajan con los ejes adicionales para la extracción de puntos de referencia. Finalmente los autores reportan un EER de $5,6\,\%$ utilizando una base de datos con imágenes recortadas y sin variaciones de pose.

Liu et al. combinan la vista frontal y la imagen trasera de la oreja extrayendo características utilizando el método de triangulación del ratio y el concepto de momentos descriptores propuesto por Tchebichef [92]. Los momentos de Tchebichef son un conjunto de funciones ortogonales basadas en polinomios discretos, fueron introducidos como un método para representación de características en 2001 [201]. La parte trasera de la oreja es descrita por una serie de líneas que son perpendiculares al eje más largo en el contorno de la oreja. Estas líneas miden el diámetro local del pabellón auricular en puntos predefinidos. El rendimiento de este enfoque combinado es de $97,5\,\%$. Si sólo se utiliza la parte frontal el rendimiento es del $95\,\%$, mientras que para las imágenes posteriores Liu et al. informan de un rendimiento del $86,3\,\%$.

Lu et al. [155], así como Yuan y Mu [149] utilizan el modelo de forma activa para extraer el contorno de la oreja. Los autores utilizaron imágenes de la oreja recortadas manualmente de 56 sujetos en diferentes poses. Un extractor de características almacena puntos seleccionados del contorno de la oreja junto con su distancia al tragus. La dimensionalidad de los vectores de características se reduce mediante el análisis de componentes principales (PCA) antes de la aplicación de un clasificador lineal. Lu et al. comparan el rendimiento donde sólo se utilizó la oreja izquierda o derecha para la identificación y también muestran que el uso de ambas orejas aumenta la precisión de 93, 3 % a 95, 1 %. En la base de datos USTB-III. Yuan y Mu reportan un rendimiento del 90 % si la rotación de la cabeza es inferior a 15 grados. Para ángulos de rotación entre 20 y 60 grados, el rendimiento cae a 80 %.

2.4.4. Clasificadores y enfoques estadísticos

Victor et al. fueron el primer grupo de investigación en transferir la idea de utilizar el espacio Eigen del reconocimiento facial al reconocimiento de orejas [43]; reportaron que el rendimiento de la oreja como característica de reconocimiento es inferior al rostro. Dichos resultados pudieron deberse al hecho que en sus experimentos Victor et al. consideraron que las orejas izquierdas y derechas debían ser simétricas. Utilizaron una oreja para entrenamiento y la otra para pruebas, lo que podría haber bajado el rendimiento del algoritmo PCA en este caso. Informaron que la precisión fue del 40 %. Con un rendimiento efectivo del 72, 2 % en la base de datos UND colección E, Chang et al. [131] alcanzaron una efectividad significativamente mayor que Víctor et al.

Alaraj et al. [157] publicaron otro estudio, donde también utilizan PCA para representación de características en el reconocimiento de orejas. En su enfoque una red neuronal multi-capa de alimentación hacia adelante fue entrenada para la clasificación de los componentes de características obtenidos con el algoritmo PCA. El resultado observado fue del 96 % de precisión; por lo tanto mejoran los resulta-

dos anteriores de Victor et al. y Chang et al.; cabe resaltar que este resultado sólo se basa en un subconjunto de una de las colecciones UND, que consta de 85 imágenes de 17 sujetos. Su conclusión fue que la oreja es tan útil para el reconocimiento como lo es el rostro.

Zhang y Mu [45] realizaron estudios sobre la eficacia de los métodos estadísticos en combinación con clasificadores. En su investigación [97] muestran que el análisis de componentes independientes (ICA) es más eficaz en la base de datos USTB-I que el PCA. Utilizaron PCA e ICA por primera vez para reducir la dimensionalidad de las imágenes de entrada y luego entrenar un SVM *(Support Vector Machine)* para clasificar la extracción de los vectores de características. Además, con la influencia de diferentes tamaños en los conjuntos de entrenamiento se midió el rendimiento: dependiendo del tamaño del conjunto de entrenamiento la efectividad del PCA varía entre el 85 % y el 94, 12 %, mientras que el rendimiento para ICA varía entre 91, 67 % y el 100 %.

Xie y Mu [252] proponen un algoritmo de incrustación lineal local mejorada *(locally linear embedding - LLE)* para reducir la dimensionalidad de las características de la oreja. LLE es una técnica para proyección de puntos de datos de alta dimensión en un sistema de coordenadas de menor dimensionalidad mientras preserva la relación entre los puntos de datos individuales. Esto requiere que los puntos de datos sean etiquetados de alguna manera, de modo que su relación esté fijada. La versión mejorada de LLE de Xie y Mu eliminó el problema mediante el uso de una función de distancia diferente. Además Xie y Mu mostraron que LLE es superior al PCA y al Kernel PCA, si los datos de entrada contienen variaciones en las poses.

Sus estudios se llevaron a cabo en la base de datos USTB III mostrando que el rendimiento regular del LLE (43 %) se mejoró significativamente con su método a 60, 75 %. Si la variación de la pose es sólo de 10 grados, el enfoque LLE mejorado alcanzó un mejor rendimiento de hasta 90 %.

En su enfoque Nanni y Lumini [147] proponen utilizar *Sequential Forward Floating Selection (SFFS)*; método iterativo estadístico para la selección de características en tareas de reconocimiento de patrones. SFFS trata de encontrar el mejor conjunto de clasificadores mediante la creación de un conjunto de reglas, que mejor se ajusten al conjunto de características. Los conjuntos se crean mediante la adición de un clasificador a la vez y evaluando su poder discriminativo con una función de aptitud predefinida. Si la nueva regla supera a la versión anterior, es agregada al final del conjunto. Los experimentos se llevaron a cabo en la base de datos UND colección E y los clasificadores individuales se fusionan mediante el uso de la regla de suma ponderada. SFFS selecciona las sub-ventanas más discriminativas que corresponden al conjunto de reglas más aptas. Nanni y Lumini reportan una tasa de reconocimiento de entre 80 % y 93 %. El EER varía de entre 6, 07 % y 4, 05 % en función del número de sub-ventanas utilizadas para el reconocimiento.

Yiuzono et al. consideran el problema de encontrar características correspondientes a las imágenes de las orejas, como un problema de optimización y aplican búsqueda genética local para resolverlo de forma iterativa [232]. Seleccionan subventanas locales con diferentes tamaños como la base para la selección genética. Yiuzono et al. [232] presentan elaborados resultados, que describen el comportamiento de la búsqueda local genética bajo diferentes parámetros, tales como diferentes métodos de selección y diferentes números de cromosomas. En una base de datos de 110 sujetos reportan una tasa de reconocimiento del 100 %.

Yaqubi et al. utilizan las características obtenidas por una combinación de detectores de bordes tolerantes a cambios en la posición y escala [169]. Este método de extracción de características se llama modelo *HMAX* y está inspirado en la corteza visual de los primates y combina características simples para entidades semánticas más complejas. Las características extraídas se clasifican con un SVM *(Support Vector Machine)* y un kNN *(k-nearest neighbors)*. El rendimiento en un pequeño conjunto de datos de 180 imágenes de orejas recortadas de 6 sujetos varía entre 62 % y 100 %, dependiendo del tipo de características básicas.

Moreno et al. implementan un extractor de características, que localiza siete puntos de referencia en la imagen de la oreja, que corresponden a los puntos más destacados de la obra de Iannarelli. Además obtienen un vector morfológico, que describe la oreja como un todo. Estas dos características se utilizan como entrada para diferentes clasificadores de red neuronal. Comparan el rendimiento de cada una de las técnicas de extracción de características individuales con diferentes métodos de fusión. La base de datos de prueba propietaria se compone de orejas recortadas manualmente con 168 imágenes de 28 sujetos. El mejor resultado fue del 93 % medido usando una red de compresión. Otras configuraciones arrojaron tasas de error de entre 16 % y 57 %.

Gutiérrez et al. [142] dividen las imágenes recortadas de la oreja en tres partes de igual tamaño. La parte superior muestra el helix, la parte media muestra la concha y la sección inferior muestra el lóbulo. Cada una de estas imágenes secundarias se descompone por la transformada wavelet y luego alimentan una red neuronal modular. En cada módulo de la red fueron utilizados diferentes integradores y funciones de aprendizaje. Los resultados de cada uno de los módulos se fusionan en el último paso para la obtención de la decisión final. Dependiendo de la combinación entre el integrador y la función de aprendizaje, los resultados varían entre 88, 4 % y 97, 47 % de rendimiento en la base de datos USTB-I.

Naseem et al. [107] proponen un algoritmo de clasificación general basado en la teoría de detección de compresión. Asumen que la mayoría de las señales son compresibles en la naturaleza y que cualquier función de compresión se traduce en una representación dispersa de esta señal. En sus experimentos en las bases de datos UND y FEUD, Naseem et al. demostraron que su método de representación dispersa es robusto frente a variaciones de pose y diferentes condiciones de iluminación. El rendimiento varió de entre 89, 13 % y 97, 83 %, dependiendo del conjunto de datos utilizado en el experimento.

2.5. Reconocimiento en imágenes 3D

En el reconocimiento de orejas 2D la variación de la pose y la posición de la cámara, también llamadas rotaciones fuera del plano, siguen siendo desafíos no resueltos. Una posible solución está en utilizar modelos 3D en lugar de fotos como referencias, porque una representación 3D del sujeto puede adaptarse a cualquier rotación, escala y traslación. Además, la información de profundidad contenida en modelos 3D se pueden utilizar para mejorar la precisión de un sistema de reconocimiento. Sin embargo, la mayoría de los sistemas de reconocimiento de orejas 3D tienden a ser costosos computacionalmente hablando. La tabla 2.5.1 resume los sistemas de reconocimiento 3D que se describen en esta sección.

Aunque el algoritmo *Iterative Closest Point (ICP)* fue diseñado originalmente para ser un enfoque para el registro de imágenes, el error de registro también puede ser utilizado como una medida de disimilitud entre dos imágenes 3D. Debido a que el algoritmo ICP se diseño para ser un algoritmo de registro, es robusto frente a todo tipo de traslaciones o rotaciones. Sin embargo tiende a detenerse demasiado pronto, ya que se queda atascado en mínimos locales. Por lo tanto requiere que los dos modelos sean pre-alineados antes de utilizar dicho algoritmo para afinar la alineación. Chen y Bhanu extraen nubes de puntos del contorno del helix exterior y el registro de estos puntos con el modelo de referencia [86]. En un enfoque posterior Chen y Bhanu [88] utilizan el algoritmo *Local Surface Patches (LSP)* para detectar los puntos característicos de la oreja, un ejemplo se presenta en la figura 2.5.1. Como el LSP contiene menos puntos que el helix exterior reduce el tiempo de procesamiento mientras que mejora el rendimiento de $93,3\,\%$ a $96,63\,\%$.

Yan y Browyer descomponen el modelo de la oreja en una matriz de elementos tridimencional *(voxels)* y extraen características de la superficie de cada uno de sus elementos. Para acelerar el proceso de alineación, cada voxel es asignado a un índice de tal manera que el ICP solamente necesite alinear el par de voxels con el mismo índice [191]. Yan y Browyer [192] proponen el uso de nubes de puntos

Tabla 2.5.1: Métodos de reconocimiento en imágenes 3D

Publicación	Resumen	Cantidad Sujetos	Imágenes	Precisión
Cadavid et al. [210]	ICP and Shape from shading	462	*NA*	95 %
Chen & Bannu [88]	Local Surface Patch	302	604	96,36 %
Chen & Bannu [86]	ICP Contour Matching	52	213	93,3 %
Liu & Zhang [91]	Slice Curve Matching	50	200	94,5 %
Islam et al. [212]	ICP with reduced meshes	415	830	93,98 %
Islam et al. [165]	Local Surface Features with ICP Matching	415	830	93,5 %
Passalis et al. [82]	Reference ear model with morphing	525	1031	94,4 %
Yan & Browyer [191]	ICP using voxels	369	738	97,3 %
Yan & Bowyer [192]	ICP using Model Points	415	1386	97,8 %
Zeng et al. [96]	Local Binary Patterns	415	830	96,39 %
Zhou et al. [128]	Surface Patch Histogram and voxelization	415	830	98,6 %, 1,6 %*EER*

para el reconocimiento de orejas en 3D, en contraste con el trabajo de Chen et al. [86] son utilizados todos los puntos del modelo segmentado. El rendimiento reportado por Yan et al. en su investigación del año 2005 [191] es del 97, 3 % y del 97, 8 % en sus posteriores trabajos del 2007 [192]. Los resultados si bien son similares no son directamente comparables porque se utilizan diferentes conjuntos de datos para la evaluación.

Cadavid et al. proponen un sistema de reconocimiento de orejas en tiempo real, que reconstruye modelos 3D a partir de imágenes CCTV 2D utilizando la forma con una técnica del sombreado [211]. Posteriormente, el modelo 3D se compara con las imágenes de referencia 3D que se almacenaron en la galería. La alineación

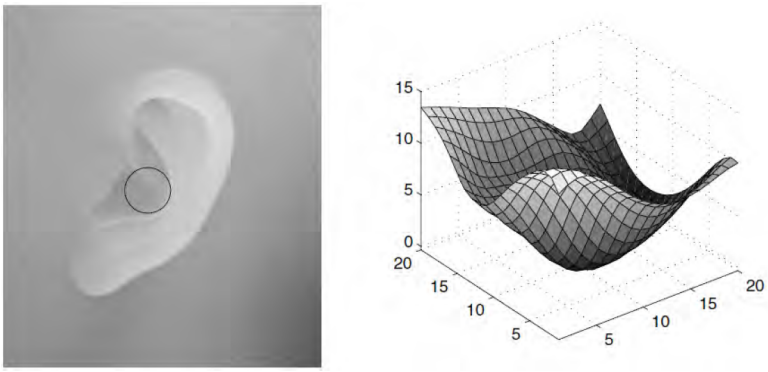


Figura 2.5.1: Ejemplo del algoritmo LSP propuesto por Islam et al [214]

de los modelos, así como el cálculo de la medida de disimilitud se lleva a cabo con el algoritmo ICP. Cadavid et al. reportan una tasa de reconocimiento del 95 % en una base de datos de 402 sujetos. Se afirma [211] que el enfoque tiene dificultades con variaciones de pose. Zhou et. al. [128] utilizan una combinación de características de histogramas locales con modelos basados en voxels. El enfoque de Zhou et. al. parece ser más rápido y con un EER de 1, 6 % también más preciso que los algoritmos de comparación propuestos por Chen y Bhanu y Yan y Browyer.

Al igual que Cadavid et al., Liu et al. reconstruyen modelos 3D a partir de vistas de la oreja 2D [92]. Sobre la base de las dos imágenes de una cámara de visión estéreo se deriva una representación 3D de la oreja. Posteriormente, el resultante 3D sirve como entrada para el algoritmo PCA. Sin embargo Liu et al. no proporcionan ningún resultado acerca de la exactitud de su sistema; suponemos que este enfoque ya no se sigue investigando debido a que ya que no publican ningún resultado.

Passalis et al. recorren un camino diferente para la comparación de modelos 3D con el fin de hacer de este proceso un enfoque adecuado para sistemas en tiempo real [82]. Se calcula un modelo de referencia de la oreja que es representativo de la oreja humana promedio.

Durante la captura, todos los modelos de referencia son deformados hasta que se ajustan al modelo promedio. Todas las traslaciones y deformaciones, que fueran necesarias para adaptarse al modelo se almacenan como características. Si una prueba de autenticación se le da al sistema, el modelo también se adapta a la oreja almacenada con el fin de obtener los datos de deformación. Posteriormente estos datos se utilizan para buscar un modelo de referencia asociado en la base de datos. En contraste con los sistemas previamente descritos, sólo una deformación tiene que ser calculada por cada intento de autenticación. Todos los demás modelos pueden calcularse antes de que se inicie el proceso de identificación en tiempo real. Este enfoque se presenta como una solución adecuada para los sistemas de reconocimiento en tiempo real, ya que toma menos de 1 milisegundo comparar dos plantillas de orejas. La tasa de reconocimiento reportada es de $94,4\,\%$. La evaluación es basada en datos no públicos, que se recogió utilizando diferentes sensores.

Heng y Zhang proponen un algoritmo de extracción de características basado en la comparación curva de secciones inspirado en los principios de la tomografía computarizada [91]. En su enfoque el modelo de oreja 3D se descompone en porciones a lo largo del eje ortogonal de la distancia más larga entre el lóbulo y la parte superior del helix. La información de la curvatura es extraída de cada porción y es almacenada en un vector de características junto con un valor de índice que indica la posición formal de la sección en el modelo 3D. Se determina la más larga secuencia común entre dos secciones curvas con índices similares. Su enfoque es solamente evaluado en un conjunto de datos no público, el cual consta de 200 imágenes de 50 sujetos. No hay información en el experimento sobre las variaciones de pose u oclusión. Heng y Zhang reportan una identificación del $94,5\,\%$ y $4,6\,\%$ EER *(Equal Error Rate)*.

Islam et al. reconectan las nubes de puntos que describen los modelos de orejas 3D en mallas y reducen iterativamente el número de secciones en ella [212]. Estas mallas simplificadas son entonces alineadas entre sí mediante el algoritmo ICP y el error de alineación se utiliza como medida de similitud. Posteriormente Islam et

al. extraen parches locales de la superficie, como se muestra en la figura 2.5.1 y los utilizan como características [214]. Para la extracción se seleccionan un número de puntos al azar a partir del modelo 3D. A continuación, se toman los puntos que están más cerca del punto semilla que del radio definido. Luego se aplica PCA para encontrar la mayoría de las características descriptivas. El algoritmo de extracción va aplicando LSP hasta que se encuentra el número deseado de características. Se evaluaron ambos enfoques usando imágenes de la base de datos UND.

La tasa de reconocimiento de Islam et al. en su trabajo del 2008 [212] es del $93,98\,\%$ y posteriormente [214] en el 2011 presentan un $93,5\,\%$. Sin embargo, ninguno de los enfoques ha sido probado con variación de pose ni diferente escalamiento.

Zheng et al. extraen un índice de cada punto del modelo 3D, el cual, es utilizado para proyectar dicho modelo en un espacio 2D [96]. Cada píxel es representado por un valor gris en la posición correspondiente en la imagen 2D. Luego se extraen las características SIFT del mapa de índices. Para cada uno de los puntos SIFT se calcula un sistema de coordenadas locales donde el eje z representa los puntos característicos en la curva normal. De ahí que los valores z de la imagen de entrada son normalizados de acuerdo a la normal de los puntos clave SIFT a los que fueron asignados. Tan pronto como se han normalizado, se transforman en una imagen de escala de grises. Como resultado, Zheng et al. obtienen una imagen a escala gris para cada una de las características SIFT seleccionadas. Luego los patrones binarios locales (LBP) se extraen para la representación de las características en cada una de las imágenes grises. Primero se comparan los índices de los puntos clave para luego utilizar Earth Mover's Distance para la comparación de los histogramas LBP de las correspondientes imágenes normalizadas en gris. Zheng et al. evaluaron su enfoque en un subconjunto de la base de datos UND en la colección J2 logrando un rendimiento de $96,39\,\%$.

2.6. Trabajos de los últimos años

Esta sección tiene por objeto proporcionar una visión general de las contribuciones de otros investigadores en el campo del reconocimiento de orejas desde mediados de 2012. Este trabajo se analiza en el contexto del estado del arte. Presenta enfoques para la segmentación, el reconocimiento de orejas en 2D y 3D.

2.6.1. Segmentación

Debido a que muchos conjuntos de datos que están a disposición del público contienen imágenes recortadas y normalizadas, la segmentación ha pasado desapercibida en muchos trabajos. La única base de datos pública, que permite experimentos de segmentación es la colección UND-J2 [193]. Lei en al. [154] mostraron que el modelo de árbol estructurado propuesto por Zhu et al. [244] también puede ser aplicado a la detección de la oreja. Además, el enfoque detecta señales fiables con un error medio entre 4,5 píxeles para UND-F y 5,5 píxeles para UND-J2. Numerosos métodos de segmentación se han propuesto en la última década con un rendimiento cercano al 100 %, lo que nos lleva a la conclusión de que la tarea de segmentación en UND-J2 puede considerarse un problema resuelto.

En los últimos años la comunidad científica comenzó a aplicar técnicas de segmentación para la detección, tal como la obra de Zhang et al. [153]. Los autores detectan la punta de la nariz y luego escanean la silueta para encontrar la región de la oreja, región con la distancia más alejada de la cámara dentro de la región de interés. Este enfoque es similar al propuesto por Yan y Bowyer [193].

Jamil et al. [176] proponen un sistema de reconocimiento que utiliza un corte parcial normalizado *(Biased normalized cut - BNC)* para segmentar la región de la oreja pre-recortada (ver figura 2.6.1). BNC es una técnica donde las regiones adyacentes de la oreja con condiciones similares de iluminación se fusionan hasta que se alcanza una condición definida. La tasa de detección obtenida es del 95 %.

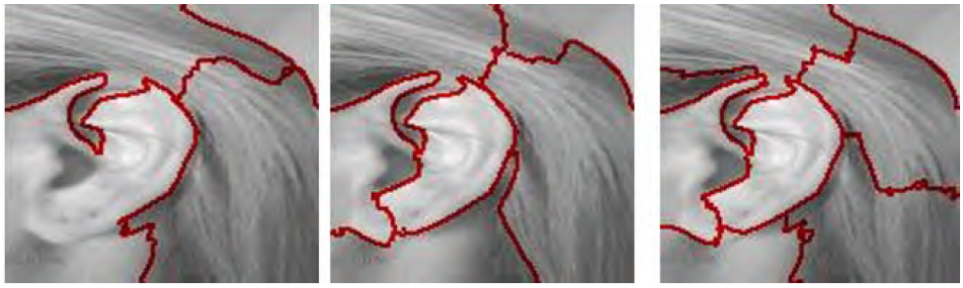


Figura 2.6.1: Ilustración de los cortes del algoritmo de normalización parcial BNC [176]

2.6.2. Reconocimiento de orejas 2D

Como ya hemos observado en el presente capítulo, hay una serie de conjuntos de datos públicos disponibles para pruebas de algoritmos de reconocimiento. El conjunto de datos más popular sigue siendo el conjunto de datos UND-J2 [191], seguido por la base de datos IITK [15].

Zhang et al. [44] han llevado a cabo una amplia investigación sobre Filtros de Gabor y su aplicación a la biometría de la oreja; introdujeron un enfoque que utiliza un filtro multi-escala de Gabor para imágenes de orejas en diferentes poses. Compararon al azar poses de la misma persona con variaciones de 0 a 60 grados. En sus resultados informaron que la tasa de reconocimiento supera el 95 %. En una publicación posterior Zhang et al. [45] extraen características de longitud definida mediante filtros Gabor y reducen la dimensionalidad con escasa representación no negativa. Similar a este enfoque, Yuan y Mu [150] también utilizan un banco de filtros de Gabor como descriptores locales, pero para la reducción de la dimensionalidad utilizan *kernel fishier discriminant analysis (KFDA)*. Ambos enfoques utilizan distancia euclídea entre dos representaciones de entidades vectoriales en el subespacio de características para realizar la comparación.

Jamil et al. [176] extraen las características de la región de la oreja segmentada también mediante filtros Log-Gabor. La salida del filtro se cuantifica con el fin de obtener una representación binaria. Estas representaciones también llamadas códigos de oreja fueron comparadas mediante la distancia de Hamming. El sistema propuesto se probo con una base de datos propia que contenía 200 imágenes de 50 sujetos.

Prakash & Gupta [221] mejoran el contraste de las imágenes de la oreja previamente recortadas y extraen características SURF de múltiples muestras. Se crea un vector de características de estas muestras que contienen todos los diversos pares de puntos de las imágenes registradas. Posteriormente se entrena un clasificador del vecino más cercano, el cual calcula las distancias euclidianas entre los puntos clave coincidentes.

Finalmente, Xu & Zheng [250] proponen un método de extracción de puntos de referencia similar al planteado por Choras [160]. Extraen bordes utilizando el detector de bordes Canny y luego montan un eje a través de la oreja que conecta el lóbulo y la helix superior. Basado en esto, se calculan proporciones del aspecto que describen la forma de la oreja y la comparan mediante distancia euclídea ponderada.

2.6.3. Reconocimiento 3D

El progreso del reconocimiento 3D aún es impulsado por algoritmos que se han desarrollado alrededor del conjunto de datos UND-J2 [192]. Este conjunto de datos se mantiene como la única base de datos pública que contiene imágenes 2D y 3D. El conjunto se compone de imágenes de perfil completo, donde sólo se muestra la cabeza de una persona. Debido a la falta de datos alternativos, la mayoría de los enfoques de reconocimiento 3D en la literatura son actualmente diseñados para imágenes de profundidad.

Zhang et al. [45] utilizan la información de profundidad inmediatamente después de la normalización de las imágenes de entrada y reduce la dimensionalidad con escasa representación. Wang y Mu [137] proponen un enfoque que automáticamente detecta los puntos claves, son detectados mediante la observación de la variación de los datos de profundidad dentro de parches locales en la imagen. Los descriptores de puntos claves se utilizan para inicializar el algoritmo ICP para el registro y la comparación de un par de imágenes de la oreja en 3D. Otro enfoque para la extracción de parches de superficie a partir de imágenes de profundidad se describe en [120]. Let et al. usan la información de la curvatura y el índice de los parches locales de la imagen para la detección automática de puntos de interés en la superficie de la oreja. Las características constan de información espacial y un factor de ponderación. En los experimentos, los descriptores de puntos de interés propuestos superan a enfoques similares, como las imágenes rotadas y previos enfoques propuestos por el mismo grupo de trabajo, denominado *Surface Patch Histogram of Indexed Shapes (SPHIS)* [130]. El rendimiento reportado es del 93 % en imágenes rotadas, frente a 97, 4 % para el método propuesto. Un enfoque similar ya fue propuesto anteriormente por Zhou & Cadavid [128]. Después de la detección de los puntos clave, Wang y Mu alinean dos imágenes usando solo los puntos claves detectados y el algoritmo ICP. El error de alineación del ICP se utiliza finalmente como la medida de similitud.

Aunque el uso de imágenes de profundidad permite la creación de sistemas computacionalmente más rápidos, cualquier información de las diferentes perspectivas de la oreja en diferentes poses se pierde. Dimov et al. [61] proponen renderizar una imagen clara de la oreja 3D segmentada en diferentes poses y luego extraer un conjunto de características de todas las imágenes renderizadas. Las tasas de rendimiento no son muy impresionantes. La idea de crear un conjunto de imágenes renderizadas parece sencillo y prometedor, pero también añade la exigencia de una estimación de pose exacta.

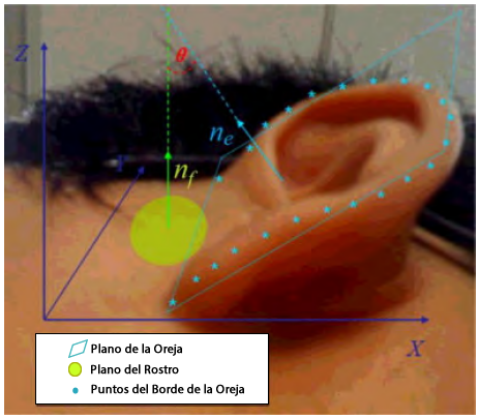


Figura 2.6.2: Concepto basado en el ángulo formado entre el rostro y la oreja [248]. De la perspectiva mostrada en la figura se puede extraer el plano de la oreja y el rostro y extraer el ángulo formado entre ellos

Cantoni et al. [234] proponen el uso de Imágenes Gaussianas Extendidas (EGI) para representar la estructura de una malla 3D sin la necesidad de crear imágenes renderizadas. Los autores recogen un conjunto de modelos de orejas de 11 diferentes sujetos y extraen un histograma EGI de longitud fija del y proporcionan una prueba de concepto de su idea; también muestran la robustez del descriptor a ruido uniforme. Debido a que el descriptor es una representación holistica de todo el objeto en 3D, su enfoque es invariante a cualquier variación en la pose. Por otro lado, el descriptor requiere que los datos de prueba y referencia sean mallas 3D, requisito que difícilmente se cumple en los sistemas de hoy en día.

Un enfoque interesante para la indexación en grandes bases de datos lo describe Liu et al en [248] quienes investigan la singularidad del ángulo entre la oreja y el plano de la cara. Un montaje experimental especial se diseño, a fin de recoger imágenes 3D de 250 sujetos desde una perspectiva especial que revele el ángulo entre el rostro y la oreja (véase la figura 2.6.2). Liu et al. proporcionan evaluaciones estadísticas de las características de unicidad y demuestran que sus vectores pueden reducir el espacio de búsqueda en grandes conjuntos de datos a tan solo 10 % de la base de datos original.

2.7. Desafíos y aplicaciones futuras

Como las publicaciones más recientes sobre el reconocimiento de orejas 2D y 3D han mostrado, la aplicación principal de esta técnica es la identificación personal en entornos sin restricciones. Esto no sólo incluye aplicaciones para vigilancia inteligente, como lo expuesto por Cadavid et al. [211], sino también la identificación forense en imágenes de circuito cerrado (CCTV) o para los sistemas de control de migración en aeropuertos y fronteras. Tradicionalmente estos campos de aplicación forman parte de sistemas de reconocimiento facial que obtienen imágenes del usuario de diferentes ángulos por lo que la oreja podría proporcionar valiosa información adicional que sirva como un segundo paso de verificación *(2-step authentication factor)*.

Sistemas multimodales que combinen reconocimiento facial y de orejas pueden servir como un medio para lograr mayor precisión sobre variaciones de pose y más robustez frente a la oclusión en entornos sin restricciones. En la mayoría de lugares públicos las cámaras de vigilancia se encuentran encima de la cabeza con el fin de captar a la mayor cantidad de personas. Además, las personas no suelen mirar directamente a la cámara, por lo que en la mayoría de los casos no existen disponibles imágenes frontales del rostro. Este hecho plantea serios problemas a los sistemas biométricos basados en reconocimiento facial. Si el rostro no es visible desde un ángulo frontal, la oreja puede volverse una característica extremadamente valiosa.

Debido a la proximidad física del rostro y la oreja, también hay muchas posibilidades de llegar a la fusión biométrica de estas dos modalidades. Se podría obtener un vector de características que incluya ambos rasgos. En el contexto de la presente investigación, hay algunos desafíos no resueltos, que deben ser abordados para futuras investigaciones en este campo.

2.7.1. Localización automática de la oreja

El hecho de que muchos de los sistemas presentados en la literatura utilicen imágenes pre-segmentadas muestra que la detección automática de orejas, especialmente en imágenes en condiciones reales, sigue siendo un problema sin resolver. Si se lograra implementar un sistema de identificación automática basado en reconocimiento de orejas se necesitaría en primera instancia de un enfoque rápido y fiable para la detección sin intervención humana. Como un primer paso para lograr este objetivo, algunos grupos de investigación han publicado colecciones de datos, que simulan variaciones típicas en entornos no controlados, tales como diferentes condiciones de iluminación, poses y oclusión. Sobre la base de estos datos, los enfoques existentes y futuros de reconocimiento deben ser probados en condiciones reales con el fin de mejorar su fiabilidad.

Por otra parte, en los últimos años los sistemas basados en imágenes 3D vienen siendo más baratos. Consecuentemente el reconocimiento 3D basado en orejas se vuelve importante y con ello la necesidad de la localización de las orejas en imágenes de profundidad o modelos 3D. Actualmente, pocos enfoques de detección en imágenes de profundidad se han publicado, aunque esto representa un primer paso para la detección en imágenes 3D.

2.7.2. Oclusión y variaciones de pose

En contraste con el rostro, la oreja se puede cubrir parcial o totalmente por el cabello o por otros artículos como pendientes, audífonos, gafas, etc. Debido a la superficie convexa del contorno externo de la oreja, partes de ella también pueden ser ocluidas si el sujeto cambia de pose. En algunas publicaciones, la robustez frente a la oclusión se aborda de manera explícita, pero no existen estudios sobre el efecto de ciertos tipos de oclusión en la tasa de reconocimiento. Una vez más, la disponibilidad de bases de datos públicas que contengan imágenes ocluidas podría fomentar el desarrollo de soluciones para cambios en la pose y algoritmos robustos

para la detección y extracción de características. Además, partiendo de la revisión de la literatura podemos afirmar que al menos en la actualidad no se encuentran estudios disponibles sobre la visibilidad de la oreja como característica biométrica de reconocimiento en ambientes públicos. Con el fin de desarrollar algoritmos para la detección y reconocimiento de la oreja, es necesaria más información sobre los tipos de oclusión más comunes en condiciones reales.

La oclusión debido a la rotación de la cabeza es otro reto pendiente en los sistemas de reconocimiento de orejas. Al igual que en el reconocimiento facial, partes de la oreja pueden llegar a ser ocluidas si se cambia de postura. Recientemente, algunos de los métodos de extracción de características que se han propuesto, han tocado este tópico, mostrando cierta robustez frente a la variación de pose en al menos en algún grado. Sin embargo, este tema no está totalmente resuelto aún. Otra posibilidad es compensando las variaciones de pose, esto podría lograrse mediante el uso de modelos 3D en lugar de fotografías 2D.

2.7.3. Escalabilidad

Actualmente las bases de datos disponibles sólo se componen con algo menos de 10.000 imágenes de orejas. La única excepción es la colección USTB-IV, que no ha sido lanzado al público todavía. En ambientes reales el tamaño de las bases de datos será significativamente más grande, lo que hace no factibles las búsquedas exhaustivas en los escenarios de identificación. Por lo tanto, no sólo la precisión, sino también la velocidad de comparación de estos sistemas de reconocimiento será interesante para futuras investigaciones.

Con el fin de hacer el reconocimiento aplicable para sistemas a gran escala, las búsquedas exhaustivas deben sustituirse por estructuras de datos adecuadas que permitan complejidad de tiempo logarítmico durante la búsqueda. Esto podría, por ejemplo, lograrse mediante la exploración de las posibilidades de organización de las plantillas de orejas en árboles de búsqueda.

2.7.4. Entendimiento simetría y envejecimiento

Debido a que el reconocimiento de orejas es uno de los últimos campos de investigación biométrica, la simetría de la oreja izquierda y derecha todavía no ha sido plenamente comprendida. Un estudio realizado por Abaza y Ross [5] indica que hay un cierto grado de simetría entre las orejas izquierda y derecha, que podría ser explotado al comparar ambas orejas. Su resultado fomenta más investigaciones sobre las restricciones de simetría existentes.

Los estudios de Iannarelli [235] indican que algunas características del contorno externo de la oreja se pueden heredar y el envejecimiento afecta ligeramente la apariencia. Ambos supuestos podrían ser confirmados en estudios más recientes, pero debido a la insuficiente falta de datos, el efecto de la herencia y el envejecimiento en la apariencia de la oreja aún no se entienden completamente.

Además, no hay aún estudios a gran escala de la relación simétrica entre las orejas izquierda y derecha. Por lo tanto otro campo interesante para la investigación futura podría ser, obtener una comprensión más profunda del efecto de la herencia y la simetría en el carácter distintivo de este modelo biométrico. Finalmente, se necesitan estudios a largo plazo sobre el efecto del tiempo en los rasgos fundamentales de las orejas a fin de obtener una mejor comprensión de la permanencia de esta característica.

Hemos presentado un estudio sobre el estado del arte en la literatura sobre la biometría de las orejas, los algoritmos existentes para su detección y reconocimiento en imagenes 2D y 3D. Hemos clasificado el gran número de enfoques basados en imágenes 2D en métodos holísticos, locales, híbridos y metodos estadísticos, discutiendo sus características y reportado su desempeño.

El reconocimiento de orejas es todavía un nuevo campo de investigación. Aunque hay un número prometedor de aproximaciones, ninguna de ellas ha sido evaluada bajo escenarios realistas que incluyan factores disruptivos como las variaciones en las poses, la oclusión y diferentes condiciones de iluminación. En los más recientes enfoques, estos factores se toman en cuenta, pero se requiere más investigación sobre esto hasta que los sistemas de reconocimiento puedan ser utilizados en la práctica. La disponibilidad de bases de datos de prueba adecuadas, que hayan sido recogidas en escenarios realistas, contribuirá aún más a la maduración de la oreja como una característica biométrica.

Hemos realizado un resumen de las bases de datos disponibles, los enfoques existentes de detección de orejas, los sistemas de reconocimiento y los problemas sin resolver para este modelo biométrico en el contexto de la vigilancia inteligente, control de fronteras, e identificación de personas que consideramos son las aplicaciones más importantes de la biometría utilizando orejas. Creemos que esta nueva característica representa una extensión valiosa para los sistemas del reconocimiento facial en el camino a la identificación automática sobre los cambios de pose.

3

Descripción del sistema

3.1. Adquisición de los datos y pre-procesamiento

La existencia de bases de datos que contengan el recorte de la oreja donde se resalten de forma perfecta sus características tubulares es limitada y escasa. No existen conjuntos de datos estándares contra los cuales se pueda contrastar el trabajo. Como resultado, existe gran dificultad en comparar adecuadamente el sistema propuesto con lo descrito en la sección 2, ya que utilizan principalmente datos privados.

En este trabajo, sin embargo, intentamos utilizar una variedad de conjuntos de datos para establecer algunos puntos de referencia sobre los que se pueden construir futuras investigaciones. Por este propósito, utilizamos un total de cuatro conjuntos de datos en nuestros experimentos; tres de estos son públicos y solamente

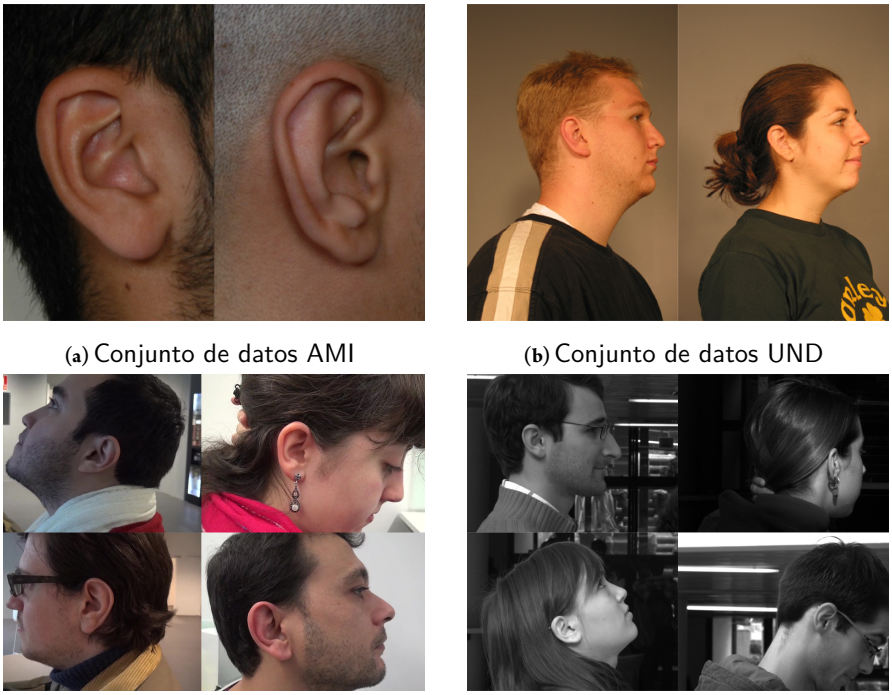
Tabla 3.1.1: Detalles del contenido de varios conjuntos de datos utilizados en este trabajo

Conjunto de datos	Tamaño	Sujetos	Imágenes por sujeto	Resolución en *pixeles*	Color Canales	Contenido	Fuente
AMI [84]	700	100	7	492 × 702	*Color*	Primer plano, ambos lados	*Foto*
UND [115, 132]	464	114	4	1200 × 1600	*Color*	Perfil, lado derecho	*Foto*
Vídeos (entrenamiento)	950	5	190	1920 × 1080	*Color*	Perfil, ambos lados	*Video*
Vídeos (prueba)	910	7	130	1920 × 1080	*Color*	Perfil, ambos lados	*Video*
UBEAR v1.0 [203] (entrenamiento)	4497	127	35	1280 × 960	*Grises*	Perfil, ambos lados, y mascaras	*Video*
UBEAR v1.1 [203] (prueba)	4624	115	40	1280 × 960	*Grises*	perfil, ambos lados	*Video*

uno es privado. Cada uno de estos conjuntos de datos tiene un conjunto de características que lo hacen particularmente útil para tareas especificas, y cada uno presenta nuevos retos. Se han utilizado dichos conjuntos para realizar una selección de experimentos reales en cada uno.

La tabla 3.1.1 proporciona una visión general del contenido de cada conjunto de datos, y la figura 3.1.1 expone algunas muestras de cada uno para demostrar cualitativamente su contenido.

El primer conjunto de datos es AMI [84]; una colección de cerca de 700 imágenes de orejas. Estas son fotografías de alta calidad con las orejas perfectamente alineadas y centradas en el marco de la imagen, también tienen alta resolución, buenas condiciones de iluminación y en buen enfoque. Este conjunto de datos es ejemplar para probar la sensibilidad de reconocimiento hacia diferentes orejas, sin embargo, debido a la naturaleza de acercamiento de las imágenes, no están realmente bien adaptadas para las tareas de localización de la oreja.



(a) Conjunto de datos AMI

(b) Conjunto de datos UND

(c) Conjunto de vídeos

(d) Conjunto de datos UBEAR

Figura 3.1.1: Muestras de cada uno de los cuatro conjuntos de datos utilizados

El segundo conjunto de datos que utilizamos es el conjunto UND [115, 132]; una colección de fotografías de perfil de varios sujetos en la cual la oreja cubre sólo una pequeña parte de la imagen. La calidad fotográfica de estas capturas es muy alta, y otra vez todo con buena y constante iluminación, y sin ninguna de las orejas ocluidas por el cabello u otros objetos. Las posturas de los sujetos varían ligeramente en relación con la cámara, pero no tanto como para introducir distracción debido a la rotación de la cabeza y pose del sujeto. Como resultado, estas imágenes son adecuadas para pruebas en la cual se específica la localización en una imagen donde la oreja representa una pequeña proporción de la totalidad del contenido, se evitan así los desafíos que representan cambios en el punto de vista *(viewpoint)* y la variación de iluminación.

El tercer conjunto de datos es obtenido de forma privada a partir de vídeos. Una colección privada de 940 imágenes compuesto de frames HD extraídos de cortas secuencias de vídeo de participantes voluntarios. Hay 14 secuencias de imágenes de siete sujetos. Cada secuencia consta de 65 fotogramas *(frames)* de una duración de aproximadamente 15 segundos extraído de un vídeo continuo. A los sujetos se les pidió que rotaran la cabeza en varias posturas naturales siguiendo movimientos suaves y continuos a lo largo de la secuencia. La iluminación y el medio ambiente son relativamente constantes a través de todos los vídeos, y a los sujetos se les pidió que removieran cualquier posible oclusión lejos de sus orejas; utilizamos este conjunto de datos para probar la sensibilidad del detector solamente en rotaciones de la cabeza en relación con la cámara, evitando al mismo tiempo los retos a iluminación variable. La mayor cantidad de imágenes por sujeto son útiles para reducir el efecto de usar una gran cantidad de formas de oreja variables para las pruebas, y de nuevo, se concentran principalmente en su pose. Una variación de este conjunto de datos fue reservado para fines de entrenamiento, el cual contenía perfiles de otros cinco participantes diferentes de los sujetos en el conjunto de datos de prueba.

El conjunto de datos final y quizás más importante que usamos es el UBEAR. Se trata de una colección bastante completa y de gran tamaño de imágenes de sujetos con una amplia gama de variaciones, que abarca múltiples dimensiones no solamente en la pose y en la rotación sino también en la iluminación, la oclusión, e incluso el foco de la cámara. Estas imágenes, por lo tanto, simulan en un muy buen grado las condiciones en entornos no cooperativos (condiciones del mundo real) donde podrían capturarse imágenes naturales para llevar a cabo tal detección y posterior reconocimiento. Dichas imágenes, aunque sin duda contienen la oreja, no intentan capturarla en perfectas condiciones, y como tal reflejan un escenario real para el proceso de prueba. Como nuestro principal interés en este trabajo es la detección de imágenes naturales no-cooperativas, este se convierte en el conjunto de datos ideal para probar el máximo potencial del sistema que proponemos.

Tabla 3.1.2: Diferencias y desafíos existentes en el conjunto de datos UBEAR

Ángulos	Abajo	Medio	Arriba	Hacia dentro	Hacia fuera
					
Exposición	Buena	Sobreexpuesta	Subexpuesta		
					
Difuminada	Si	No			
					
Genero	Masculino	Femenino			
					
Oclusión	Si	No			
					

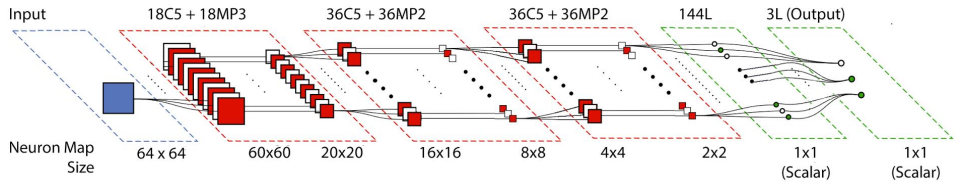


Figura 3.2.1: Arquitectura de la red neuronal convolucional utilizada

La tabla 3.1.2 es una muestra del conjunto UBEAR donde se aprecia la complejidad del contenido, el cual que se asemeja a condiciones del mundo real. También es importante tener en cuenta que el conjunto de datos UBEAR dispone de dos versiones, las cuales consisten en imágenes únicas en ambos conjuntos. La primera de estas versiones, llamada 1.0, incluye una máscara sobre el fondo delineando la ubicación exacta de la oreja en cada imagen; cómo se describirá más adelante, este aporte fue importante para nuestro procedimiento de entrenamiento. La versión 1.1 no incluye tal máscara y es sin embargo reservada para pruebas y experimentación.

3.2. Detección

3.2.1. Red neuronal convolutional

La red neuronal convolucional *(Convolutional Neural Network - CNN)* utilizada es un estándar con algunas modificaciones en su arquitectura que ayudan a manejar particularmente este escenario. La arquitectura de red utilizada se representa visualmente en la figura 3.2.1.

Dado que el objetivo final de este sistema es la detección y reconocimiento de orejas en tiempo real a partir de secuencias de vídeo e imágenes 2D, se requiere un sistema que se ejecute rápidamente. Por lo tanto, necesitamos una arquitectura altamente optimizada que sigue siendo útil para reconocer los datos de entrena-

miento. Las clases objetivo que buscamos reconocer con las redes neuronales son tan sólo tres: (i) oreja izquierda, (ii) oreja derecha, y (iii) el fondo; de aquí en adelante referidos por sus abreviaturas LE, RE, y BG respectivamente. La intra-clase o coeficiente de correlación entre clases en el conjunto es, en definitiva, muy similar con todas las orejas siguiendo un conjunto parecido de rasgos distintivos. Como resultado, la red no necesita aprender una cantidad ilimitada de características diferentes, como sería el caso en las grandes CNN's que se utilizan para el reconocimiento de imagen a gran escala. En cambio, un modesto tamaño, con un número limitado de neuronas resulta ser suficiente para llevar a cabo la tarea de detección.

Dada la naturaleza de las imágenes, se decidió que un tamaño óptimo de entrada para la red sería de 64×64, ya que una oreja de ese tamaño es lo suficientemente grande para incorporar la información necesaria para identificación y definir adecuadamente la forma y sus componentes internos, y a la vez no es demasiado grande para requerir una convolución de gran tamaño o agrupaciones de capas demasiado complejas para analizarlas correctamente.

Finalmente, necesitamos recorrer la imagen en su totalidad para identificar los objetos dentro, esto se puede lograr construyendo un recuadro que se mueve a lo largo de la imagen cómo si de una ventana se tratase; el algoritmo de ventanas deslizantes *(sliding-window)* debe incluir un factor acumulativo máximo que no debe ser demasiado grande. Esto asegura que el tamaño de cada paso correspondiente en el mapa de salida no excederá el tamaño permitido. Por lo tanto, se decidió un máximo de tres capas convolucionales y sus correspondientes capas de agrupamiento. Conocer estas tres restricciones, para la entrada y salida, y la máxima cantidad de capas, a través de un proceso de ensayo y error iterativo, derivó al final en la arquitectura definida por la ecuación 3.1.

$$CNN = 18C_5 : MP_3 + 36C_5 : MP_2 + 36C_5 : MP_2 + 144L + 3L \qquad (3.1)$$

Donde, *CNN* es la red neuronal, *C* indica que es una capa (*layer*) convolucional, *MP Max Pooling Layer* y *L* una capa lineal, los números previos a las letras representan la cantidad de neuronas; de este diseño se produce entonces un tamaño de paso entre ventanas mínimo de $3 \times 2 \times 2 = 12$, que es bastante eficaz para el propósito de detección, ya que permite analizar en intervalos de no más de 12 píxeles. La elección de este parámetro también es bastante beneficioso, ya que el número 12 tiene muchos factores que más tarde ayudarán a alinear las ventanas de detección en escalas múltiples.

3.2.2. Inferencia triple red profunda

El entrenamiento de una única red neuronal y esperar que sea suficiente para distinguir adecuadamente las orejas del ruido que representa el fondo *(background)* en imágenes obtenidas en un ambiente no controlado es un gran salto de fe.

En la práctica, una red neuronal de este tipo es capaz de reconocer apropiadamente un gran porcentaje de los objetos representados con la forma de la oreja. Por lo tanto, cuando estas redes se prueban frente a un conjunto de orejas recortadas y preparadas específicamente para la tarea de reconocimiento su desempeño es destacable, es decir proporcionan tasas de acierto elevadas. Sin embargo, son propensas a cometer muchos errores cuando se le presentan imágenes con un fondo con excesivo ruido. De lo anterior que se haya decidido entrenar la red con una clase fondo *(background - BG)* para ayudarle a aprender la diferencia entre una oreja y una con ruido, pero no importa cómo se prepare el entrenamiento para esta clase; una CNN siempre será propensa a falsas detecciones simplemente debido a la funcionalidad interna de las redes neuronales. Siempre habrá patrones o combinación de características que se pueden encontrar fácilmente en imágenes naturales que activarán la neurona equivocada y, por tanto, producir una gran tasa de falsos positivos. La tabla 3.2.1 describe este efecto con más detalle. Una sola CNN detectará frecuentemente la oreja correctamente, tanto en imágenes de cerca como en la AMI (99,70 %), y también en imágenes complejas del conjunto de

Tabla 3.2.1: Desempeño comparado de 1-CNN contra 3-CNN

Conjunto de datos	Algoritmo	Verdaderos positivos	Falsos positivos	Falsos negativos	Orejas detectadas (%)	Métrica F1 (%)
AMI	**1-CNN**	698	0	2	99.70	99.86
	3-CNN	693	0	7	99.00	99.50
UBEAR	**1-CNN**	4326	11935	280	93.90	41.46
	3-CNN	3814	605	661	82.80	85.77

datos UBEAR $(93,90\,\%)$; sin embargo, esta métrica ignora el efecto de falsos positivos. La métrica F1 es útil para descubrir la gran disparidad que se produce en realidad: mientras que en el conjunto de datos AMI, el valor F1 sigue siendo alto $(99,86\,\%)$, en el conjunto de datos UBEAR cae abismalmente $(41,46\,\%)$ debido a la gran cantidad de falsos positivos introducidos. La métrica F1 *(F1 Score)* tiene en cuenta la precisión p y la recuperación *(recall)* r. Es la media armónica de estas dos métricas y se calcula según la formula 3.2.

$$F1 = 2\frac{(p \times r)}{(p + r)} \tag{3.2}$$

Este problema generalmente se puede resolver creando múltiples clasificadores. Todos analizan la misma entrada de datos, y sus diferentes salidas se combinan para crear un resultado final cuya precisión suele ser mayor que la de cualquier simple clasificador ejecutándose por sí solo [68].

Aplicamos una variación de esta idea: no procesamos todos los clasificadores en el conjunto con los mismos datos exactos de entrada, sino que presentamos diferentes datos a cada componente del conjunto. Por lo tanto cada uno de los cla-

Figura 3.2.2: Las tres escalas utilizadas en el set de entrenamiento

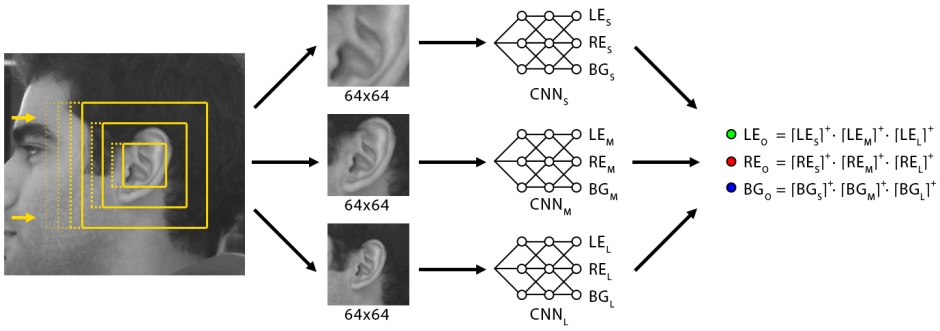


Figura 3.2.3: Flujo de datos en el proceso de inferencia con 3-CNN's

sificadores deben ser entrenados para especializarse en los tipos de datos que serán presentados. Las diferentes entradas de datos se construyen cuidadosamente para cada uno, es decir, datos específicos a cada componente según su propia especialización.

La idea principal entonces es alimentar a tres redes neuronales tres imágenes distintas, correspondiendo cada una a la misma región que se está analizando a diferentes escalas. La figura 3.2.2 representa las escalas que recibe el clasificador. Etiquetamos adecuadamente cada una de las redes utilizadas como: S *(small)*, M *(medium)* y L *(large)*.

El objetivo de las tres escalas es principalmente el entrenamiento especializado de las redes para los propóósitos específicos de: (i) reconocer las características tubulares de la oreja interna, (ii) enmarcar las coordenadas correctas de la oreja, e (iii) inferir el contexto de la oreja dentro de una región circundante. Entrenar una red con una sola de estas escalas provocaría que se especializara particularmente en la data ignorando elementos distintivos de su entorno, y a la vez la red sería ajena a otros datos de imagen en ambiente natural con estructura similar pero no realmente pertenecientes a una oreja verdadera, y así llevarlo a producir una gran cantidad de falsos positivos, lo que terminaría por afectar la precisión de detección. Sin embargo, las tres redes trabajando juntas como un comité de clasificadores produce un resultado mucho más robusto que desemboca en un clasificador más resistente ante el ruido, concluyendo en una mayor certeza, ya que se requerirá la activación de las tres redes para obtener una respuesta, simplemente integrando la información contextual en el sistema.

Cada una de las tres redes neuronales produce tres valores de salida, que corresponden a la probabilidad de que cada clase objetivo haya sido percibida en esta entrada de la red. Denotamos los valores de salida como O_A^K, donde $A \in \{S, M, L\}$ representa el índice de red indicado por su tamaño, y $K \in \{LE, RE, BG\}$ representa el índice de la clase de salida de cada red, para cada uno de los posibles resultados detectados. Cada una de estas salidas estará en el rango $[-1, +1]$ al haber sido las redes neuronales entrenadas con esos valores ideales.

Para combinar los resultados de las tres redes como un conjunto unificado, filtramos cada clase de salida con los valores correspondientes en las tres redes después de que cada uno haya sido rectificado linealmente. Las salidas finales del conjunto están definidos por la formula 3.3.

$$
\begin{aligned}
O_F^{LE} &= \lceil O_S^{LE} \rceil^+ \cdot \lceil O_M^{LE} \rceil^+ \cdot \lceil O_L^{LE} \rceil^+ \\
O_F^{RE} &= \lceil O_S^{RE} \rceil^+ \cdot \lceil O_M^{RE} \rceil^+ \cdot \lceil O_L^{RE} \rceil^+ \\
O_F^{BG} &= \lceil O_S^{BG} \rceil^+ \cdot \lceil O_M^{BG} \rceil^+ \cdot \lceil O_L^{BG} \rceil^+
\end{aligned}
\tag{3.3}
$$

Donde $\lceil x \rceil^+ \equiv max(0, x)$ es una operación de rectificación lineal. Pasando sólo los valores positivos de cada salida evitamos la interferencia de múltiples valores negativos, cualquiera de los cuales tiene entonces el efecto de poner en cero la neurona final. La figura 3.2.3 representa el proceso visualmente. El efecto neto de este proceso, entonces, es que las tres redes funcionen en fila, donde sólo sobrevivirán las regiones donde las tres redes neuronales estén de acuerdo. Además, la salida final será pesada por la red individual, por lo tanto, las regiones en las que las tres neuronas de salida tengan alta probabilidad superará a las regiones donde la distribución de salida es más dispareja.

La idea tras tener tres redes neuronales separadas ejecutándose a diferentes escalas es sencillamente la de obtener el menor numero de falsos positivos posible. Las redes neuronales suelen contener bastante ruido por naturaleza, y los resultados que entregan usualmente requieren ser fortalecidos con información adicional que incremente la certeza de sus predicciones. Comparar las predicciones a múltiples escalas es un enfoque común en los algoritmos de detección y es precisamente lo que hemos deseado implementar al utilizar tres escalas distintas.

La red neuronal opera de tal forma que sólo puede analizar pequeñas porciones de imagen de tamaño 64×64. Este inconveniente se solventa utilizando algoritmos como el de movimiento de ventanas (*sliding window*), donde las imágenes son divididas en regiones de 64×64 recorriendo la imagen completa y pasando cada región a la red neuronal. Adicionalmente, las diferentes escalas de la misma región son procesadas simultáneamente recortando regiones un poco más grandes que

64×64 y posteriormente reajustando de vuelta al tamaño de entrada. En general, esto es una operación intensiva que consume muchos recursos, aunque existen optimizaciones dentro del algoritmo sliding window que pueden acelerar su ejecución como es el uso de mapas compartidos.

Al utilizar tres escalas, el sistema se ve forzado a utilizar un enfoque de deslizamiento de ventanas múltiple que alinea correctamente las regiones a recortar y redefine el tamaño y sobrepone las regiones más grandes sobre la región inicial en tres escalas consecutivas. El efecto es que la imagen se vuelve una densa rejilla, donde, terminamos con múltiples ventanas recortadas alrededor de un punto, y cada subconjunto de tres escalas consecutivas se utiliza como entrada de nuestro conjunto de redes neuronales.

Cada red individual analiza el marco en su escala correspondiente, produciendo un conjunto de cuadros delimitados por las regiones para las que se entrenó la red en particular. Cuando una activación se produce en las tres redes en todas las escalas, entonces se puede deducir que una oreja existe en ese lugar, y por lo tanto este resultado tiene una mayor precisión de lo que cualquiera de las tres redes haría por sí sola.

El resultado final se obtiene combinando los resultados de estas tres detecciones. El método de combinación es una simple multiplicación de cada salida correspondiente a través de las tres redes, de tal manera que sólo los resultados donde todas las redes están de acuerdo se mantienen. Antes de la multiplicación los valores son rectificados, con el fin de truncar los valores negativos y evitar los negativos dobles que desencadenan falsos positivos. La formula 3.4 muestra la salida resultante de la multiplicación del resultado de las CNN.

$$Y_O^i = \lceil Y_S^i \rceil^+ \cdot \lceil Y_M^i \rceil^+ \cdot \lceil Y_L^i \rceil^+ \tag{3.4}$$

La observación de un resultado final por encima de un umbral determinado de forma empírica, 0,5 en las pruebas, indica la existencia de una oreja en la ventana actual que se está analizando. El cuadro delimitador final se toma de la escala del detector de nivel medio; las coordenadas se dibujan para ilustrar las detecciones finales. El enfoque combinado produce resultados mucho más limpios y más precisos de lo que cualquiera de las redes individuales podría proporcionar.

3.2.3. Datos de entrenamiento

Cómo el sistema comprenderá tres redes neuronales individuales, los datos de formación deberán recogerse de conformidad con los requisitos para cada una de las redes individuales. Se discutió previamente que cada red analizará esencialmente tres recortes de tamaño diferentes para cada región, por lo que los datos pueden ser preparados simultáneamente simplemente recortando las orejas con distintas regiones extendidas a fin de generar los datos de los otros dos tamaños.

Los conjuntos de datos existentes que consisten en fotografías de orejas segmentadas son muy escasos y pequeños en volumen. Crear grandes cantidades de capturas de entrenamiento, requiere mucho trabajo en la manipulación de imágenes. El conjunto de datos UBEAR fue particularmente útil, como se describe en la sección 3.1; incluye para cada una de sus imágenes una máscara que facilita el proceso de extraccción de la oreja. Esta máscara describe la localización exacta de la oreja en cada imagen, particularmente útil para identificar y crear los bordes delimitadores en los que esta contenida cada oreja. No todas las máscaras podían utilizarse ya que muchas eran extremadamente borrosas y no eran apropiadas para el entrenamiento. Al final, se utilizaron aproximadamente 3.000 imágenes de este conjunto para entrenamiento.

Además, complementamos los datos con muestras adicionales que se obtuvieron manualmente de fotogramas de vídeo. Con la información adicional el conjunto de entrenamiento quedó formado por 4.000 imágenes. Para aumentar aún más el tamaño del conjunto de entrenamiento, los datos se aumentaron de dos formas: (i) las imágenes fueron modificadas al azar añadiendo pequeñas traslaciones, rotaciones y re-escalando el contenido; (ii) las imágenes fueron rotadas horizontalmente y la imagen resultante fue asignada al conjunto de datos de orejas opuestas. Este proceso artificial representó un incremento de diez veces el tamaño de los datos de entrenamiento. Lo que hizo un total de aproximadamente 40.000 imágenes, es decir 20.000 para cada oreja.

Posteriormente se procesaron las imágenes para cada oreja en tres conjuntos separados por escala 3-CNN: S, M, y L. Esto se hizo simplemente recortando y cambiando el tamaño de cada muestra apropiadamente. El proceso se repitió para ambos lados, derecha e izquierda, produciendo así seis colecciones de imágenes para oreja izquierda y derecha, en cada una de las tres escalas. Finalmente, también se creó un conjunto de datos con ruido de fondo, del mismo tamaño que los otros, que consitió en recortes aleatoriamente formados de fotografías de la base de datos flickr y de regiones no-orejas de los conjuntos de entrenamiento UBEAR y los vídeos obtenidos previamente. El resultado son siete colecciones distintas para propósitos de entrenamiento, cada una de las cuales consta de aproximadamente 20.000 imágenes. La figura 3.2.4 muestra un ejemplo del resultado.

3.2.4. Entrenamiento de la red neuronal

La versión final del clasificador fue entrenada con la colección de las tres escalas descritas anteriormente. Cada una de las tres redes utilizó un conjunto de entrenamiento de 3 clases compilado a partir de orejas izquierda y derecha en la escala correspondiente, y una copia de la colección de imágenes de fondo *(background)*.

Figura 3.2.4: Un pequeño subconjunto de cada una de los bases de datos utilizados para entrenamiento. De arriba hacía abajo: izquierda-pequeña, izquierda-mediana, izquierda-grande, derecha-pequeña, derecha-mediana, derecha-grande, fondo (background)

La estructura de las tres redes es exactamente la misma, descrita en la sección 3.2.1. La entrada consta de un sólo canal en escala de grises con la imagen redimensionada a un cuadrado de tamaño 64×64. Las imágenes de entrada se pasan a través de una etapa de pre-procesamiento que consiste en un *Spatial Contrastive Normalization (SCN)*, que ayuda a mejorar los bordes de la imagen y redistribuir el valor medio y el rango de datos, algo que mejora el entrenamiento de CNN's.

El conjunto de entrada resultante pasa por la primera capa de extracción, la cual consiste de 18 neuronas convolucionales, una ReLU nonlinearity *(rectified linear unit)* y tres capas de agrupamiento *(pooling layers)*. A continuación los datos pasan por dos capas similares de extracción, cada capa de 36 neuronas, para posteriormente pasar por una capa lineal de 576 neuronas y finalmente la capa de salida con tan sólo tres neuronas que corresponde a cada clase del conjunto de entrenamiento fondo, oreja-izquierda y oreja-derecha.

Tabla 3.2.2: Matriz de confusión final de la data de entrenamiento

Clasificado / Clase correcta	Oreja izquierda	Oreja derecha	Fondo	Total clase	Exactitud (%)
Oreja izquierda	16040	56	88	16184	99,11
Oreja derecha	46	16064	74	16184	99,26
Fondo	63	194	15927	16184	98,41
			Total	48552	98,93

Tabla 3.2.3: Matriz de confusión final de los datos de prueba

Clasificado / Clase correcta	Oreja izquierda	Oreja derecha	Fondo	Total clase	Exactitud (%)
Oreja izquierda	3964	34	49	4047	97,95
Oreja derecha	14	4002	31	4047	98,89
Fondo	8	42	3997	4047	98,77
			Total	12141	98,54

Cada red ha sido entrenada con sus correspondientes datos es decir los recortes pequeños, medianos y grandes. Un enfoque SGD (*Stochastic Gradient Descent*) estándar fue utilizado para el entrenamiento y se ejecutó en aproximadamente 24 iteraciones hasta que ya no se obtenían mejoras en el test-fold de los datos. Los valores ideales para cada uno de los resultados de las distintas etiquetas se asignaron en el rango $[-1, +1]$, donde las etiquetas activas son valores positivos, y las etiquetas inactivas los valores negativos.

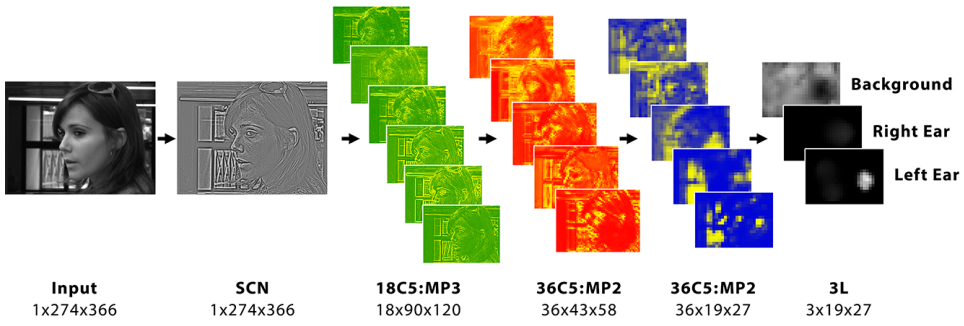


Figura 3.2.5: Ejecución del mapa compartido de una de las CNN's sobre una imagen ejemplo de entrada

Esta distribución fue elegida, a diferencia del rango más tradicional $[0, 1]$, para mejorar el rendimiento de las tres CNN's como se explica en la sección 3.2.2. Todos los conjuntos de datos se dividen en conjuntos de entrenamiento y prueba, en una ratio $80\,\%/20\,\%$ de acuerdo a las prácticas estándar de aprendizaje en machine learning. El resultado final de entrenamiento sobre estos dos conjuntos se resumen en las tablas 3.3.1 y 3.2.3.

3.2.5. Detección en ejecución

El funcionamiento en tiempo de ejecución de la red se realizó a través de mapas compartidos como entrada de las CNN's. Esto permitió ejecutar un método optimizado para inferir la detección y predicción en una imagen completa de una manera mucho más eficiente que el enfoque tradicional de ventana deslizante *(sliding-window)*.

El proceso requiere que la imagen de entrada se prepare primero como una escala multi-piramidal. Esto es; ser capaz de detectar las orejas en todos los tamaños posibles relativos al cuadrado de la imagen para poder llevar a cabo correctamente la detección, independientemente de la distancia relativa del sujeto a la cámara.

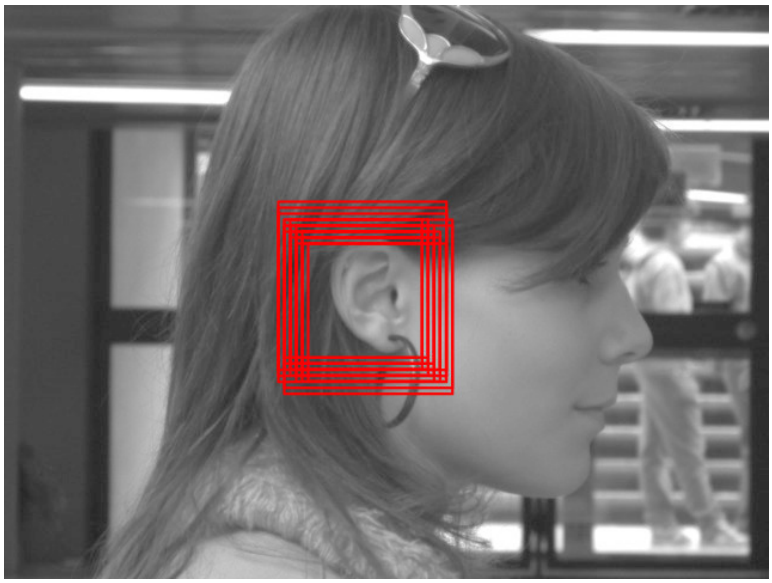


Figura 3.2.6: Muestra de múltiples detecciones superpuestas moldeadas como ventanas individuales de detección en una imagen de entrada

Cada uno de estos niveles de la pirámide se dará a cada una de las tres redes para ser analizado independientemente. Cada red, por lo tanto, crea tres mapas de salida por nivel correspondiente a cada una de las clases objetivo entrenadas: LE, RE y BG. La figura 3.2.5 representa la ejecución de mapa compartido de una de las redes para una nivel particular de la pirámide de tamaño 274×366.

Cada píxel en cada uno de estos mapas de salida corresponde a la clase predictiva; una ventana cuya localización se puede rastrear hasta la entrada de acuerdo con la alineación del mapa compartido y la configuración de su posición. La figura 3.2.6 muestra cómo las ventanas pueden reconstruirse a partir de estos mapas compartidos y corresponden precisamente a las múltiples detecciones que un sistema tradicional basado en el algoritmo de ventana deslizante produciría.

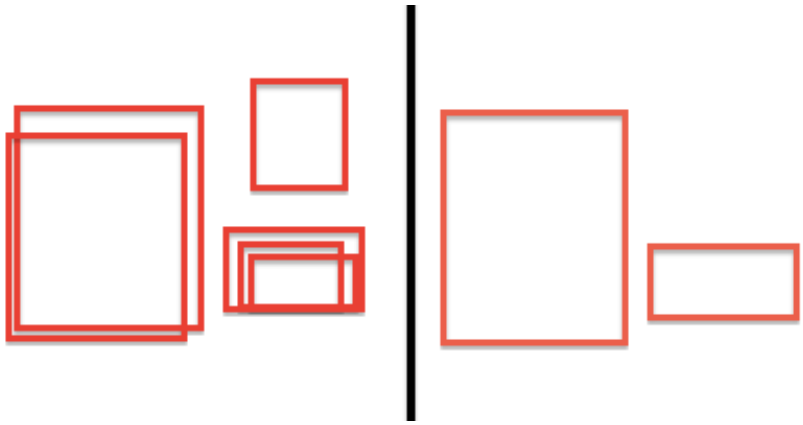


Figura 3.2.7: Muestra de cómo el algoritmo de particionamiento y agrupamiento limpia múltiples ventanas de detección sobrepuestas

Con el fin de colapsar estas detecciones múltiples en un único resultado final se utiliza un algoritmo de partición basado en estructuras de datos Disjoint-set. El algoritmo Disjoint-set es muy similar al grupo de rectángulos y funciones de partición de OpenCV [49], pero personalizado de alguna manera particular. Este algoritmo permite el agrupamiento de ventanas colocadas y escalonadas de manera parecida ya que todas ellas pertenecen a una sola detección del objeto. La figura 3.2.7 muestra un diagrama de como el algoritmo de agrupación se comporta en varios grupos de ventanas.

Esta es una práctica muy común tomada como un procedimiento de limpieza post-procesamiento en muchas tareas de visión por ordenador. Para este trabajo en particular, sin embargo, se ha creado una regla de agrupación especial para ponderar la tolerancia de la agrupación. Para cada una de las dos clases positivas, LE y RE, se realiza el procedimiento que mostramos a continuación.

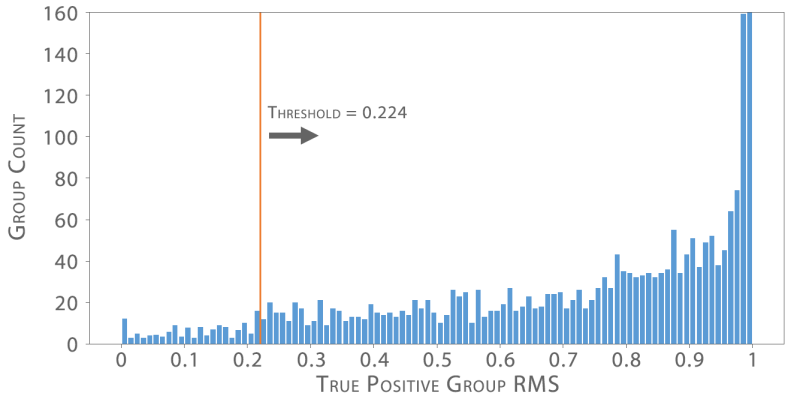
Cada ventana i tiene un valor asignado correspondiente al valor predictivo de salida de la red neuronal, denotado por O_i; esta ventana pesa su propio valor elevándolo al cuadrado. Por lo tanto, las ventanas con poco valor de predicción tie-

nen en general un nivel de importancia reducido, mientras que las ventanas con un valor de salida mayor, al inicio del proceso, pueden mantener su posición en la agrupación. Para un agrupamiento potencial j compuesto por múltiples N ventanas, cada una con un valor de salida con peso O_i^2, el valor de salida final G_j para el grupo es entonces dado por la ecuación 3.5.

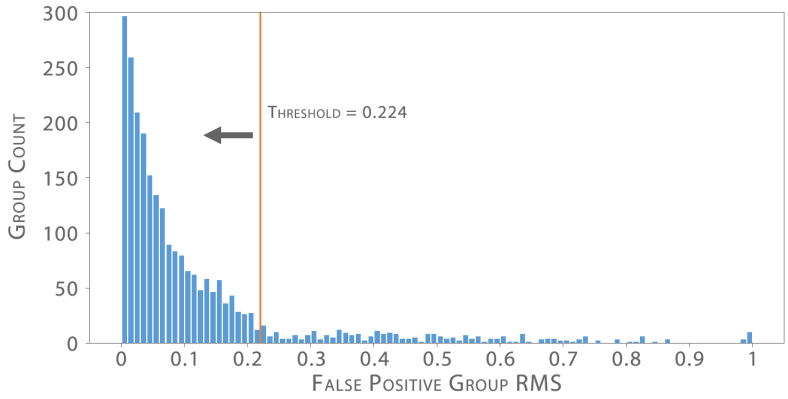
$$G_j = \sqrt{\frac{\sum_i^N O_i^2}{N}} \tag{3.5}$$

El valor de salida G_j corresponde a una raíz cuadrada media *(root mean square - RMS)* de todos los valores de salida de la ventana que se componen en ese grupo. El resultado final es que el proceso favorece a los *clusters* que se componen de ventanas con valores grandes de confianza donde además las ventanas con baja confianza, como en el caso de falsos positivos, terminan con un valor más bajo, permitiendo su exclusión. Como cada clúster tiene un único valor numérico asignado a él, de acuerdo con su importancia en general, se puede por tanto, con base en el umbral definido, rechazar aquellos con un nivel bajo de confianza.

Con el fin de obtener un umbral adecuado se realizó un experimento con todo el conjunto de datos de prueba del conjunto UBEAR. Todos los clusters generados en este proceso se clasificaron manualmente como verdaderos positivos o falsos positivos. La figura 3.2.8a muestra la distribución de los valores de salida clasificados como verdaderos positivos, y la figura 3.2.8b muestra los grupos clasificados como falsos positivos.



(a) Sensibilidad al ubmbral de verdaderos positivos



(b) Sensibilidad al umbral de falsos positivos

Figura 3.2.8: Resultados de las CNN's para los grupos TP y FP

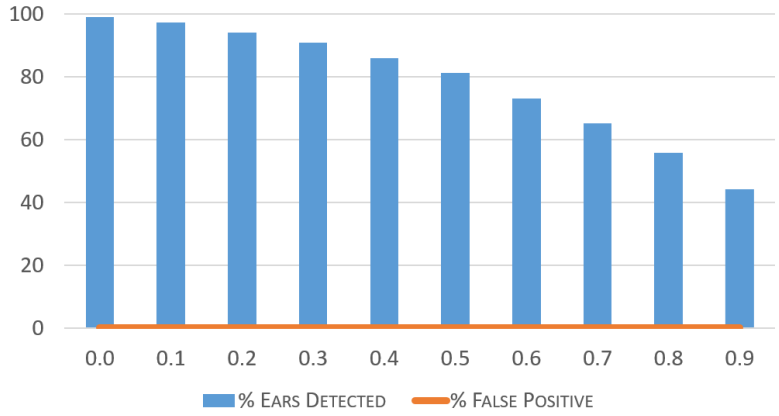
Después de analizar estas distribuciones puede verse que el valor umbral elegido de 0, 224 los separa de forma optima, donde se puede lograr un equilibrio al rechazar el mayor número de falsos positivos, manteniendo al mismo tiempo tantos verdaderos positivos como sea posible por encima del umbral. Un resumen de este proceso se muestra a continuación:

for all $Z \in \{PyramidScales\}$ **do**
    for all $A \in \{L, M, S\}$ **do**
        $O_A^{LE}, O_A^{RE}, O_A^{BG} \leftarrow SharedMap(Image_Z, Network_A)$
    end for

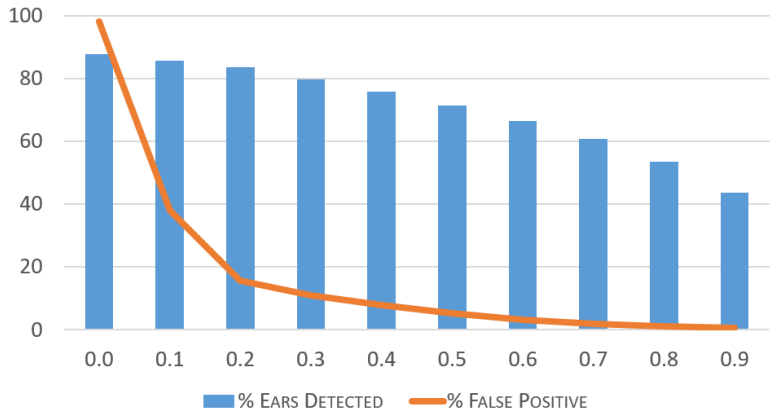
    for all $K \in \{LE, RE, BG\}$ **do**
        $O_{F,Z}^K \leftarrow Ensemble(O_S^K, O_M^K, O_L^K)$
    end for
end for

for all $K \in \{LE, RE\}$ **do**
    $G^K \leftarrow Group(O_{F,Z}^K)$

    if $G^K > Threshold$ **then**
        $Keep(G^K)$
    else
        $Discard(G^K)$
    end if
end for



(a) Conjunto de datos AMI



(b) Conjunto de datos UBEAR

Figura 3.2.9: Sensibilidad del umbral en la detección de orejas

El umbral correcto debe ser cuidadosamente seleccionado dependiendo del tipo de datos que se estén analizando. En el caso de la base de datos AMI, donde las imágenes de orejas ya están preparadas y recortadas, el sistema no detecta falsos positivos en absoluto, y por lo tanto el valor umbral no afecta la tasa de falsos positivos de ninguna manera. En este caso, un umbral muy bajo o cero puede ser elegido para maximizar la cantidad de orejas correctamente detectadas. En la figura 3.2.9a mostramos los resultados donde la tasa de exactitud del umbral representa diferentes cantidades.

En el caso de las imágenes naturales en entornos no cooperativos como el conjunto de datos UBEAR, el efecto de los falsos positivos es mucho más importante cómo puede verse en la figura 3.2.9b, donde pequeñas variaciones en el valor del umbral conducen a una drástica caída en la tasa de falsos positivos, mientras que no afecta la precision significativamente en la exactitud de las orejas detectadas.

3.2.6. Experimentos

Se realizaron múltiples experimentos con los diversos conjuntos de datos para evaluar la precisión del sistema en diferentes escenarios.

3.2.6.1. Metodología de prueba

Para todos los escenarios, los experimentos se llevaron a cabo con el método de 3-CNN propuesto anteriormente. Para contrastar los resultados, se realizaron los mismos ensayos con un clasificador básico en cascada *(Haar Cascade Classifier)* entrenados sobre datos similares implementados en OpenCV [49] y ejecutado con una configuración similar de ventana deslizante y el post-procesamiento con el mismo algoritmo de agrupación de ventanas.

Para todos los casos, los resultados reportados se describen a continuación:

- **Verdaderos positivos (TP):** detecciones que encuentran con éxito una oreja dentro de la imagen.
- **Falsos positivos (FP):** detecciones que clasifican mal la oreja detectada, o que detectan erróneamente ruido en la imagen que no corresponde a la oreja actual.
- **Falsos negativos (FN):** orejas en una imagen que no pudieron ser detectadas, o cuyo valor de confianza dentro del grupo de detección está debajo del umbral definido.
- **Verdaderos negativos (TN):** este valor describiría generalmente la tasa a la cual el ruido no-orejas es ignorado con éxito por el clasificador. Sin embargo, en este caso, aumentaría la exactitud de clasificación de forma innecesaria. Se evitó por ello grabar esta información a propósito de modo que los resultados representen solamente la verdadera naturaleza del clasificador, es decir, las orejas correctamente clasificadas.

Las métricas de rendimiento reportadas para todos los casos son: (i) la precisión que mide la exactitud del clasificador; (ii) el *recall* o recuperación que mide su integridad; y (iii) la métrica F1 que proporciona un equilibrio entre precisión y recuperación y es por tanto, una comparación más objetiva del rendimiento de los dos clasificadores. Además, también se presenta la tasa de exactitud tradicional, con el fin de proporcionar una métrica básica de rendimiento.

3.2.6.2. Comparación con el estado del arte

Debido a la variada naturaleza del estado del arte en este campo, es muy difícil realizar un estudio comparativo sobre el rendimiento de nuestro método propuesto con los métodos existentes en la literatura. En parte, esto se debe a que no hay

Tabla 3.2.4: Resultados de las pruebas sobre el dataset de vídeos

		Haar				3-CNN			
	Tamaño subconjunto	Precisión (%)	Recuperación (%)	Exactitud (%)	Métrica F1 (%)	Precisión (%)	Recuperación (%)	Exactitud (%)	Métrica F1 (%)
Alineada	470	97,60	97,60	95,32	97,60	99, 57	99, 79	99, 36	99, 68
Hacia arriba	162	100	69,75	69.75	82.18	95, 95	91., 3	87, 65	93, 42
Hacia abajo	284	98,77	57,09	56,69	72,36	94, 83	95, 19	90, 49	95, 01
Oreja izquierda	455	97,85	71,21	70,11	82,43	97, 07	97, 29	94, 51	97, 18
Oreja derecha	461	98,53	88,57	87,42	93,29	97, 98	96, 46	94, 58	97, 21
Conjunto completo	**916**	**99,05**	**80,07**	**79,45**	**88,55**	**97,59**	**96,95**	**94,68**	**97,27**

un conjunto de datos estándar con el cual todos estos algoritmos puedan ser comparados, y cada método examinado hasta ahora en la sección 2 tiende a utilizar sus propios datos privados. Del mismo modo, probar los métodos existentes sobre los mismos datos que hemos utilizado resulta complicado ya que la mayoría de las implementaciones expuestas permanecen privadas y su código fuente no está disponible para su implementación.

Por lo tanto, sólo podemos contribuir a las tablas 2.3.1 y 2.3.2 con nuestros resultados en los conjuntos de datos UND y AMI, que son imágenes de cualidades similares a los datos utilizados en esos estudios, y consisten en imágenes hechas para este propósito. En el caso de imágenes recortadas de cerca, tales como las imágenes de AMI, nuestro sistema 3-CNN alcanza una precisión de 99 % y una métrica F1 de 99, 50 %. En imágenes donde la oreja cubre el fotograma completo, como las de UND, donde la localización también juega un papel interesante, nuestro sistema alcanza una precisión de 95, 25 % y una métrica F1 de 97, 57 %. Más detalles de estos resultados se encuentran en la tabla 3.2.7.

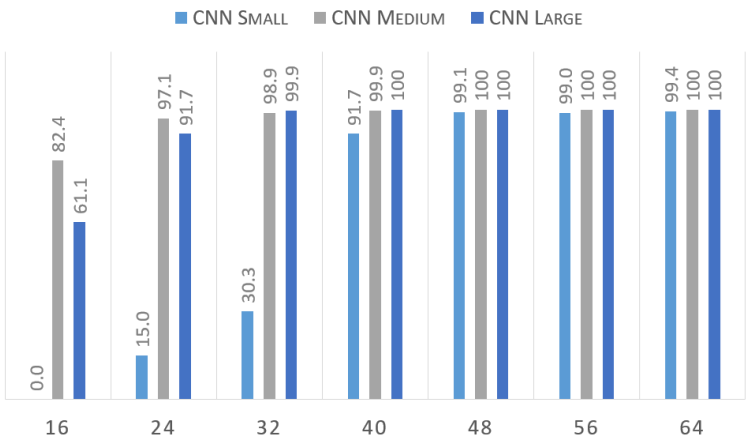
3.2.6.3. Análisis de imágenes por segundo en vídeo

Adicionalmente, también probamos la precisión de detección en imágenes por segundo en vídeo individuales. Se llevó a cabo un experimento con el conjunto de datos de vídeo como se describe en la sección 3.1. El propósito de esta prueba es asegurar que ambas orejas puedan ser correctamente clasificadas izquierda y derecha, mientras se trabaja con poses variables de la cabeza. Los resultados de esta prueba son presentados en la table 3.2.4, donde se puede ver que nuestro sistema supera en gran medida al clasificador básico Haar en esta particular tarea.

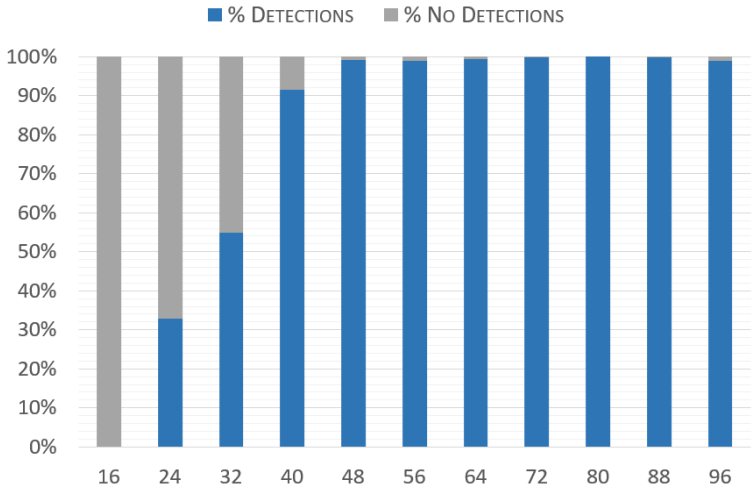
La importancia de esta prueba radica en la capacidad de detectar continuamente la misma oreja en una secuencia de imágenes en movimiento, independientemente de la orientación de la cabeza. La alta tasa de detección asegura que la oreja se detecte constantemente durante la mayoría de la duración de cada vídeo, exceptuando algunos frames en los que la detección puede fallar. Sin embargo, transcurridos unos cuantos frames, la oreja vuelve a ser encontrada y la detección continúa como se espera. Este resultado permitiría por lo tanto crear un mecanismo de seguimiento para ser implementado con éxito en secuencias de vídeo.

3.2.6.4. Resolución de la imagen

Detectar orejas de sujetos ubicados a gran distancia de la cámara suele ser problemático. Para medir cuantitativamente el rendimiento del sistema en casos en los que el tamaño relativo de la imagen es muy pequeño, se realizaron varias pruebas con el conjunto de datos AMI con los orejas previamente redimensionadas a diferentes escalas, que van desde 16×16 hasta 96×96. Los resultados de ambos sistemas combinados 3-CNN así como el de las individuales S, M y L están representadas en las figuras 3.2.10a y 3.2.10b. Esto demuestra que incluso las orejas que se encuentran en escalas mucho más bajas que el tamaño de entrada de 64×64 pueden ser detectadas con éxito, aunque con una menor tasa dependiendo del tamaño.



(a) Red convolutional individual (1-CNN)



(b) Red convolucional triple (3-CNN)

Figura 3.2.10: Resolución de imágenes y sensibilidad ante el tamaño de la oreja

Tabla 3.2.5: Desafíos existentes en imágenes en condiciones no ideales

Ángulos	Abajo	Medio	Arriba	Frontal	Lateral
					
Exposición	Normal	Iluminación alta	Iluminación baja	Oclusión	Sin oclusión
					

La figura 3.2.10a, en particular, explica la caída de de la tasa de rendimiento en escalas más pequeñas. La red convolucional S es la primera que falla en escalas decrecientes, como podría esperarse debido a la naturaleza de los datos que analiza esta red. Mientras tanto, las otras dos CNN's continúan detectando con suficiente precisión, incluso en las escalas más pequeñas. Se podría decir, sin duda, que un sistema fuera de la escala S podría hacerlo mejor para este propósito en particular.

A pesar de la disminución de la tasa de exactitud la CNN de escala pequeña ha demostrado ser esencial para la diferenciación del ruido, y como tal, su integración con el resto de redes hace del inconveniente antes descrito un efecto aceptable, al menos en el escenario de 3-CNN's. El bajo desempeño de esta red es la principal razón detrás de construir 3-CNN con el fin de superar la dificultad de detectar orejas en escalas variables.

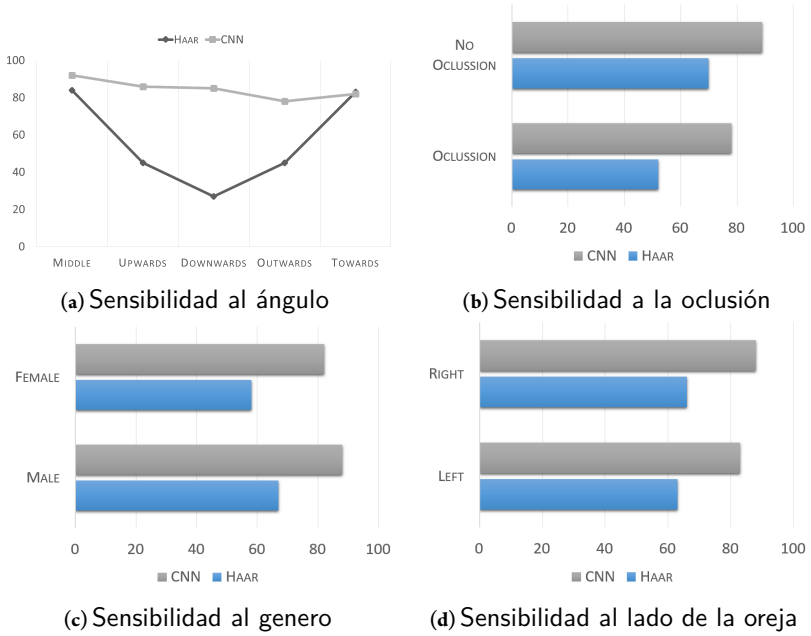
Tabla 3.2.6: Desempeño del sistema propuesto 3-CNN's frente a la clasificación HAAR en el conjunto UBEAR

		Haar				3-CNN			
	Tamaño subconjunto	Precisión (%)	Recuperación (%)	Exactitud (%)	Métrica F1 (%)	Precisión (%)	Recuperación (%)	Exactitud (%)	Métrica F1 (%)
Alineada	1392	95,90	74,31	72,03	83,74	89,89	93,92	84,95	91,86
Hacia arriba	813	85,87	30,42	28,97	44,93	87,47	84,94	75,73	86,19
Hacia abajo	784	88,65	16,23	15,90	27,44	85,68	84,57	74,09	85,12
Hacia afuera	789	89,96	30,34	29,35	45,37	85,92	71,26	63,81	77,91
Hacia adentro	829	95,10	73,24	70,57	82,75	79,70	84,40	69,47	81,99
Masculina	3403	94,17	51,58	49,99	66,65	86,10	87,84	76,93	86,96
Femenina	1204	91,06	42,47	40,77	57,92	86,99	77,83	69,71	82,15
Oreja izquierda	2289	93,49	47,33	45,82	62,84	83,64	83,11	71,48	83,37
Oreja derecha	2318	93,42	51,07	49,30	66,04	88,97	87,32	78,79	88,14
Oclusión	1491	89,70	36,49	35,03	51,88	85,01	71,63	63,60	77,75
No Oclusión	3116	94,70	55,24	53,59	69,78	86,79	91,53	80,34	89,10
Conjunto completo	**4607**	**93,45**	**49,22**	**47,58**	**64,48**	**86,31**	**85,23**	**75,08**	**85,77**

3.2.6.5. Imágenes no cooperativas en entorno natural

Los enfoques tradicionales de visión por ordenador usualmente requieren que la oreja esté perfectamente alineada, o por lo menos en la misma ubicación, dadas las restricciones que el analizar imágenes del mundo real impone; un ejemplo de estas condiciones no ideales se observa en la tabla 3.2.5. Debido a la capacidad de las CNN's para aprender múltiples representaciones del mismo objeto, y teniendo en cuenta la variedad de poses utilizadas en el entrenamiento de los datos, el sistema final es capaz de detectar orejas en ángulos muy diferentes con respecto a la cámara.

El conjunto de datos UBEAR contiene etiquetas para cada imagen que facilita su partición de acuerdo con la posición relativa del sujeto en relación con la cámara. Las pruebas se realizaron sobre todo el conjunto de datos y los resultados se dividieron según el ángulo de la mirada del sujeto. Estos resultados se representan en figura 3.2.11 y se resumen en la tabla 3.2.6.



(a) Sensibilidad al ángulo

(b) Sensibilidad a la oclusión

(c) Sensibilidad al genero

(d) Sensibilidad al lado de la oreja

Figura 3.2.11: Desempeño del sistema propuesto 3-CNN versus la clasificación HAAR en el conjunto UBEAR

Los resultados proyectados demuestran que el algoritmo propuesto supera la clasificación mediante HAAR. Sin embargo, la verdadera trascendencia es que los clasificadores en cascada, no muy diferentes a la mayoría de enfoques tradicionales, es altamente dependiente de la perspectiva, si su rendimiento cae en gran medida es debido a que el ángulo varía al mover la cabeza el sujeto. Mientras tanto, nuestro sistema 3-CNN's mantiene una clasificación muy similar y un rendimiento estable independientemente del ángulo en el que se presenta la oreja gracias a la capacidad de las redes profundas de interpretar múltiples versiones del mismo objeto.

Adicionalmente se pueden usar las etiquetas del conjunto UBEAR para dividir los datos en grupos, por ejemplo podríamos agrupar por la oblicuidad de la oreja. Cómo era de esperar, el sistema funciona casi igual para cada lado izquierda y derecha. Las pequeñas diferencias en los resultados se deben a una variación aleatoria en las imágenes, y no a que el clasificador favorezca un lado.

Se realizaron pruebas del sistema con imágenes que contienen cierto grado de oclusión comparando el resultado con lo obtenido con orejas no ocluidas. Para este estudio, agrupamos manualmente las fotografías basados en una decisión subjetiva de cuales imágenes podrían considerarse ocluidas. Esto se debe a que los grados de oclusión pueden variar desde mechones de pelo o un pequeño pendiente, accesorios muy grandes o secciones completas de pelo que cubren más de la mitad de la oreja.

La decisión final del umbral de oclusión fue hecha para marcar solamente las orejas que tuvieran su contorno cubierto en al menos un 25 % del área total que forma la oreja. El resultado aproximadamente un tercio de las imágenes fueron marcadas como ocluidas. Como era de esperar, el sistema 3-CNN funciona mejor cuando no hay oclusión. Sin embargo, vale la pena señalar que incluso al analizar las orejas ocluidas, el sistema 3-CNN supera al clasificador HAAR incluso cuando se analizan las imágenes claramente visibles, y no ocluidas.

Además, analizando la literatura de los sistemas existentes de detección, tales como los descritos en las tablas 2.3.1 y 2.3.2, es obvio que la mayoría aparentemente tienen altas tasas de exactitud reportadas por supuesto en imágenes donde la oreja esta bien definida, fallando drásticamente cuando la oreja está ocluida de alguna manera; este comportamiento se puede evaluar especialmente en aquellos sistemas que basan su análisis en la geometría de la oreja utilizando la detección de la forma tabular o hélice de una oreja.

Finalmente, se realizó un estudio final sobre la sensibilidad de género del detector. Los clasificadores no son necesariamente sensibles a las diferentes formas de la oreja masculina y femenina. Sin embargo, se puede ver una disparidad visible, simplemente debido al hecho que las orejas femeninas son mucho más propensas a ser ocluidas por el cabello más largo o el uso de más accesorios, tales como

Figura 3.2.12: Detecciones en imágenes particularmente difíciles del conjunto de datos UBEAR, incluyendo orientaciones extremas de la cabeza y oclusión

pendientes grandes. Por lo tanto, los resultados de sensibilidad de género se asemejan mucho a los de la sensibilidad a oclusión que hemos venido tratando. La figura 3.2.12 presenta algunas muestras seleccionadas de la aplicación del 3-CNN y su detección particularmente en imágenes desafiantes, debido a la oclusión o a la perspectiva con la que se ha capturado la imagen.

3.2.7. Resultados

En conclusión, la tabla 3.2.7 enumera un resumen de todos los resultados totales en los cuatro conjuntos de datos mientras comparamos nuestro sistema 3-CNN con el conocido algoritmo de clasificación Haar Cascade. Como se ha podido observar, el sistema basado en redes convolucionales siempre supera el HAAR en todos los conjuntos, en una cantidad que oscila entre el 10 % y el 29 % utilizando la métrica F1.

Tabla 3.2.7: Resumen de los resultados sobre los 4 conjuntos de datos contratando el clasificador HAAR versus el algoritmo 3-CNN

Conjunto	Algoritmo		Positivo	Negativo	Precisión (%)	Recuperación (%)	Exactitud (%)	Métrica F1 (%)
UND [115, 132]	3-CNN	**Positivo**	461	20	95,84	99,35	95,25	97,57
		Negativo	3	0				
	Haar	**Positivo**	270	7	97,47	58,44	57,57	73,07
		Negativo	192	0				
Vídeos	3-CNN	**Positivo**	890	22	97,59	96,95	94,68	97,27
		Negativo	28	0				
	Haar	**Positivo**	727	7	99,05	80,07	77,47	87,31
		Negativo	181	0				
AMI [84]	3-CNN	**Positivo**	693	0	100	99,00	99,00	99,50
		Negativo	7	0				
	Haar	**Positivo**	382	7	98,20	55,12	54,57	70,61
		Negativo	311	0				
UBEAR [203]	3-CNN	**Positivo**	3814	605	86,31	85,23	75,08	85,77
		Negativo	661	0				
	Haar	**Positivo**	2227	156	93,45	49,22	47,58	64,48
		Negativo	2298	0				

En el conjunto de datos UBEAR el mejor desempeño de la red profunda es particularmente cierto, debido a que el clasificador de Haar es incapaz de modelar la mayor variedad de representaciones internas necesarias para clasificar adecuadamente las imágenes en ese conjunto de datos. En la figura 3.2.13 se puede notar que el sistema tiene cifras de rendimiento estables en los primeros tres conjuntos de datos, los cuales consisten en fotografías perfectas de la oreja. La precision sólo disminuye ligeramente cuando se procesan imágenes naturales debido a los desafíos ya descritos. Esto contrasta con el algoritmo Haar, que tiene resultados tremendamente dispares, demostrando su gran dependencia a las condiciones particulares de un conjunto u otro.

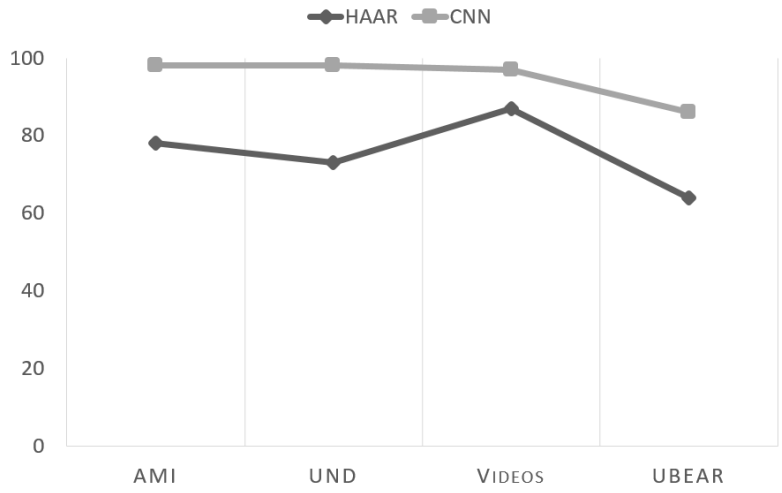


Figura 3.2.13: Resultados del sistema 3-CNN's comparado con el clasificador HAAR sobre varios datasets

3.2.8. Conclusiones

En lo que va del presente capitulo hemos presentado un nuevo enfoque para la detección de orejas basado en redes neuronales convolucionales. Las CNN's trabajan utilizando percepción de imagen y forma, que tiende a ser un enfoque mucho más robusto comparado con los sistemas tradicionales de visión artificial que se basan en características manualmente definidas para cada tarea específica; este es el caso especialmente cuando se prueba con diferentes perspectivas y que además presentan oclusión, situación muy común en imágenes naturales.

Todos los sistemas anteriormente propuestos generalmente fallan de una manera u otra: algunos requieren que la oreja esté bien alineada y otros que la oreja este visible completamente o son muy sensibles a la iluminación y requieren imágenes tomadas en las mismas condiciones que los datos de entrenamiento y fallan cuando las imágenes no están completamente enfocadas o cuando la oreja no representa una buena área de la fotografía.

Hasta ahora, no hemos visto un sistema tan robusto que sea capaz de detectar orejas bajo todas las condiciones posibles en imágenes naturales, y por ende los resultados aquí expuestos alientan a presentar esta nueva alternativa. Por supuesto, nuestro sistema todavía tiene características importantes que debemos abordar, principalmente disminuir la tasa de falsos positivos. No obstante, los resultados son muy alentadores, y contar con un detector tan robusto es el paso previo ineludible hacia la construcción de un sistema de reconocimiento de orejas el cual se describe en la siguiente sección.

Otras futuras líneas de investigación incluirán también la aplicación de este sistema de una manera aún más optimizada para poder desplegarlo en dispositivos móviles o para aplicaciones prácticas. Finalmente, es importante señalar que aunque este trabajo se enfocó hacia la detección de orejas, presenta una estructura de reconocimiento de objetos de extremo a extremo que se puede adaptar a otras tareas de visión que requieren un tipo comparable de clasificación ejecutada sobre imágenes naturales en tiempo real.

3.3. Reconocimiento

Una vez realizado el proceso de detección procedimos a identificar la oreja del individuo para lo cual se debía comparar con las imágenes de la base de datos completa. Los primeros resultados haciendo uso de una red neuronal convolucional no fueron los deseados obteniendo cantidad de posibles candidatos, muchos de los cuales debían considerarse como falsos positivos. La gran cantidad de errores reportados nos obligó a replantear el diseño del reconocimiento, definiendo una serie de pasos de pre-procesamiento y aplicando distintos enfoques para llevar a cabo el objetivo.

3.3.1. Alineamiento

El primer gran paso fue el de remodelar nuestra base de datos colocando todas las orejas alineadas de tal forma que el lóbulo siempre estuviera en el centro de la imagen; esta transformación facilitaría posteriormente la aplicación de enfoques geométricos para dar solución al problema planteado.

El proceso de alineamiento no sólo se realizó con las imágenes almacenadas en la base de datos, sino que también se aplicó a cada detección realizada con la red neuronal convolucional expuesta en la sección 3.2. El alineamiento representa el paso previo obligado que ha de realizarse con el propósito de (i) reducir la cantidad de posibles candidatos y (ii) reducir el tiempo de procesamiento de los algoritmos de reconocimiento haciendo más eficiente su desempeño.

3.3.2. Pre-procesamiento

Con la oreja detectada y debidamente alineada aplicamos dos algoritmos: (i) la transformación de imágenes basada en rayos (*IRT*) y (ii) la distancia Hausdorff. La figura 3.3.1 muestra los pasos a seguir donde se aplican ambas técnicas: se enmascaran las imágenes para posteriormente convertirse a su representación binaria, aplicar la distancia Hausdorff y determinar el número de posibles candidatos.

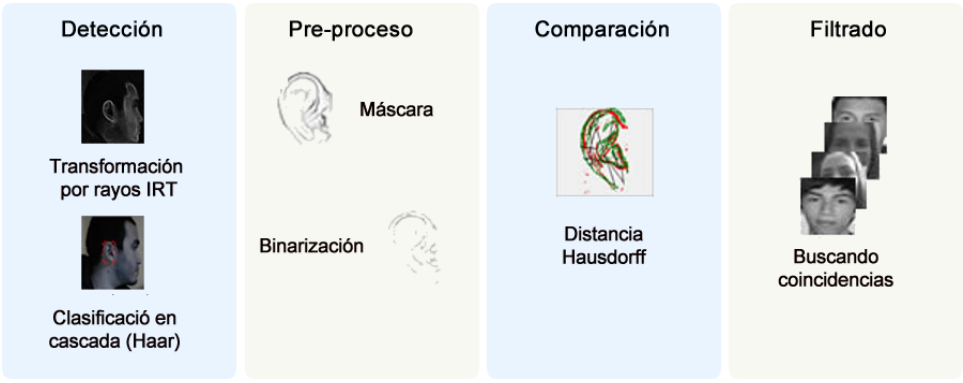


Figura 3.3.1: Pre-procesamiento de la imagen

3.3.3. Transformación de imagen basada en rayos

La transformación de imágenes basada en rayos (*IRT - Image Ray Transform*) es una técnica novedosa para mejorar las características estructurales dentro de las imágenes. Emplea los principios de la óptica de rayos para resaltar las características tubulares y circulares dentro de las imágenes. Aunque la transformación se basa en una analogía a la luz, el método puede ser modificado para mejorar la detección de funciones o mejorar el rendimiento mientras se mantiene la ventaja de la formulación analógica [11].

El IRT funciona disparando rayos a través de una imagen. La imagen transformada se genera registrando el curso de cada rayo, lo que da como resultado características estructurales mejoradas. Analizamos la imagen como una matriz de bloques de vidrio bidimensionales, cada uno de los cuales representa un píxel en la imagen. El índice de refracción de cada bloque se deriva de la intensidad de su correspondiente píxel, a través de la relación lineal con un parámetro que controla el índice de refracción máximo (n_{max}). Para una intensidad dada, i, el índice de refracción, n_i, es el detallado en la formula 3.6.

$$n_i = 1 + (\frac{i}{255})(n_{max} - 1) \tag{3.6}$$

Ese índice de refracción también podría derivarse de la relación exponencial de la intensidad del píxel con un parámetro k que controla el crecimiento de la intensidad, según lo presentado en la ecuación 3.7.

$$n_i = e^{\frac{i}{k}} \tag{3.7}$$

Dentro del arreglo de bloques, se crea un rayo con una posición aleatoria (x, y) y una dirección inicial (d). Se calcula un vector unitario del rayo V a partir de d. El rayo se inicializa en la posición del vector p y en la iteración t como se muestra en las formulas 3.8 y 3.9.

$$p^{<0>} = (x, y)^T \tag{3.8}$$

$$p^t = p^{t-1} + V \tag{3.9}$$

Cuando el rayo entra en un bloque por primera vez, la matriz acumulada A de ese bloque se incrementa en 1. Cada rayo individual incrementa un píxel sólo una vez, para evitar pequeños ciclos en el camino del rayo causando que de forma repetitiva se esté incrementando el valor de un grupo de píxeles en detrimento de la imagen completa; situación que normalmente es causada por ruido en la imagen y no por las características estructurales que se desean resaltar. Cuando el rayo cruza los limites de un bloque se calcula una nueva dirección.

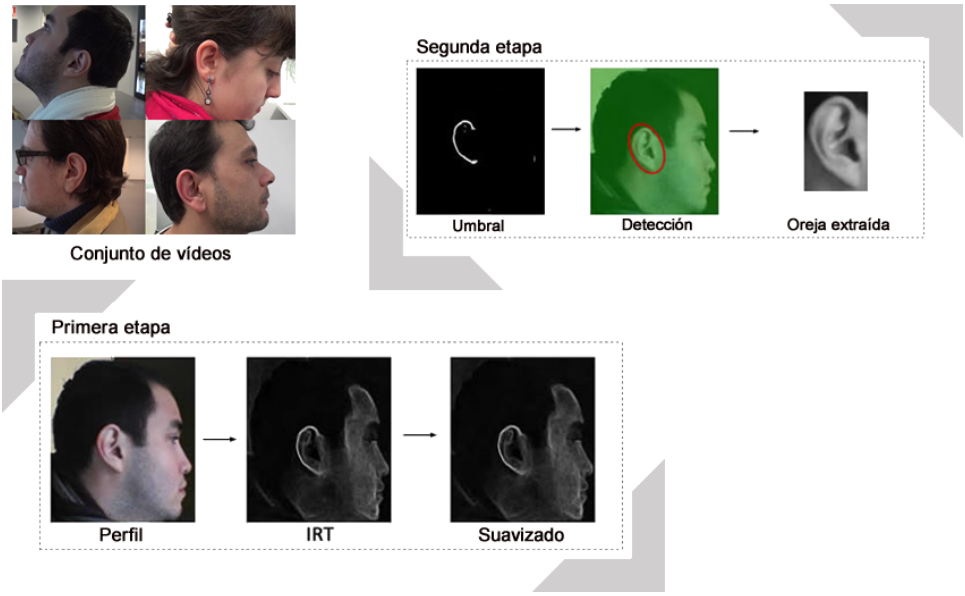


Figura 3.3.2: Pasos principales para identificar la oreja utilizando el algoritmo de transformación de imágenes por rayos

El método se basa fuertemente en una analogía con rayos de luz, hemos introducido un cambio con el fin de mejorar el resultado de la transformación, intentando evitar comprometer el algoritmo. La función que se utiliza para calcular la nueva dirección del rayo no se establece cuando se cruzan los limites horizontales y verticales de cada bloque, como sería lo esperado, sino basado en la curva normal de la dirección del borde encontrado por el algoritmo de pre-procesamiento de imágenes Sobel ya que es más representativo de la información dentro de la imagen.

La figura 3.3.2 muestra como a partir de las capturas realizadas se pre-procesa la imagen a fin de resaltar las características tubulares de la oreja. Una vez que dichos elementos se encuentran resaltados se procede a enmascarar la oreja para intentar reducir el fondo y ruido que pueda presentarse. Este paso se observa en la figura 3.3.1, la binarización genera una imagen en fondo blanco de la oreja y sólo de la oreja con sus componentes visiblemente mejorados, para finalmente proceder a la aplicación de la distancia Hausdorff.

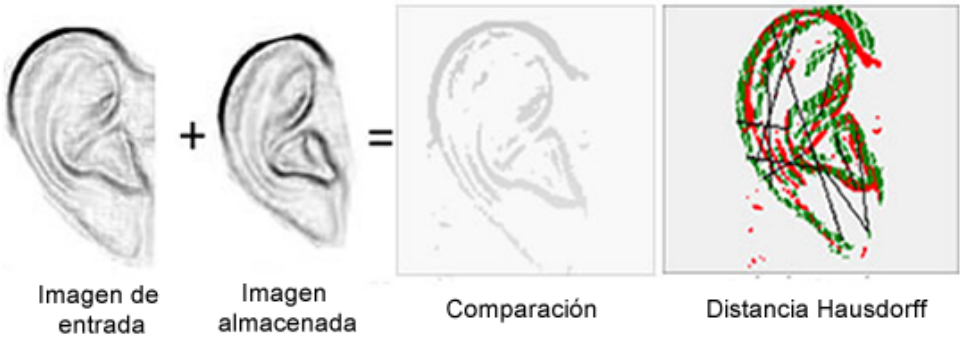


Figura 3.3.3: Distancia Hausdorff

3.3.4. Distancia Hausdorff

La medida de distancia Hausdorff utilizada en el presente documento se basa en la suposición de que las regiones de la oreja tienen diferentes valores de importancia, donde características como la helix, anti-helix, tragus, anti-tragus, concha, lobulo y el contorno de la oreja juegan el papel más importante en el proceso de reconocimiento. El algoritmo aplicado se basa en los expuesto por Kwan et al[143].

El algoritmo básicamente opera realizando una comparación de los mapas de bordes obtenidos previamente con el IRT. La ventaja de usar bordes para hacer coincidir dos objetos, es que dicha representación es robusta al cambio en iluminación. La figura 3.3.3 representa un ejemplo de la distancia de Hausdorff tratando de juntar dos imágenes. El algoritmo trata de calcular la distancia entre los puntos a fin de poder elegir un grupo de imágenes de nuestra base de datos; esta tarea funciona como un filtro eligiendo y descartando algunas imágenes para fortalecer el sistema de clasificación.

El procedimiento consiste en eliminar el fondo de la imagen tal como se realizó en el pre-procesamiento. Después de enmascarar la imagen, procedemos a obtener los bordes utilizando el filtro Canny y Sobel, aplicamos el algoritmo IRT, la imagen se invierte para operar con un fondo blanco, luego la oreja se binariza, pro-

cedimiento similar se aplica a cada imagen almacenada en la base de datos. Con los objetos obtenidos comparamos píxeles para obtener la similitud entre figuras geométricas; esto usando la distancia Hausdorff, resultando en una colección de valores que contienen la distancia de la imagen de entrada con respecto a cada elemento de la base de datos.

El objeto puede presentarse como el candidato o los candidatos que tienen la menor distancia relativa; de no superar el valor umbral mínimo, se considera como un problema sin resolver. En el sistema desarrollado, el algoritmo de Hausdorff se presenta como una tarea de pre-procesamiento complementario para aumentar el rendimiento de las tres redes neuronales convolucionales. Dichas redes se compararon para validar si una de ellas identifica al usuario, y de ser así, se acepta que la imagen pertenece dicho usuario.

3.3.5. Red convolucional

La red en la que se basa nuestro sistema es una CNN estándar compuesta por capas convolucionales alternas y de agrupación máxima para la etapa de extracción de características, y una o más capas lineales para la etapa de clasificación final. La figura 3.2.1 describe la estructura de capas de la red, y es la arquitectura de referencia utilizada para describir los conceptos presentados. La primera capa consiste en una o más neuronas que contienen los datos de la imagen a analizar, normalmente compuestos por un sólo canal de escala de grises de la imagen entrante.

La red neuronal convolucional utilizada es bastante estándar con algunas consideraciones hechas en su arquitectura que ayudan a nuestro caso de uso particular. Dado que nuestro propósito final para este sistema es el reconocimiento de las orejas en vídeo, se requiere un sistema que se ejecute rápidamente. Por lo tanto, se necesita una arquitectura altamente optimizada que siga siendo útil para reconocer los datos entrenados.

Las clases objetivo que buscamos reconocer con la red neuronal son sólo dos: (i) usuario, (ii) otra información como por ejemplo el fondo *(background)* de aquí se hace referencia por sus abreviaturas US y BG. Los datos intra-clase son, en última instancia, muy similares con todas las orejas siguiendo un conjunto similar de rasgos distintivos. Como resultado, la red no necesita aprender una cantidad ilimitada de características diferentes como sería el caso en las CNN de gran tamaño que se utilizan para el reconocimiento de objetos en general. En cambio, una red neuronal de tamaño modesto, con un número pequeño de neuronas resulta ser suficiente para aprender el tipo de datos requeridos para esta tarea. El enfoque seguido en nuestra estrategia es uno-a-muchos, significa que el sistema crea una red neuronal por individuo. Esto es aplicable sólo en un escenario controlado con pocos usuarios como la investigación que estamos llevando a cabo, donde nuestro objetivo es identificar si es posible utilizar este tipo de redes combinándolas con otros algoritmos en experimentos independientes.

Dada la naturaleza de las imágenes, se decidió que un tamaño de entrada óptimo para la red sería de 64×64, ya que el recorte de la oreja en ese tamaño es lo suficientemente grande para llevar suficiente información y definir correctamente la forma de una oreja, a la vez no son excesivas en tamaño para requerir núcleos de convolución enormes o un numero ingente de capas de agrupación. Adicionalmente nos permite mantener un estándar con la CNN utilizada en la detección. Decidimos por tanto un máximo de tres capas de convolución más sus respectivas capas de agrupación.

Finalmente, es importante resaltar que en imágenes de gran tamaño dónde la oreja representa un pequeño recuadro se utilizó el método de ventanas deslizantes para realizar el recorte perfecto de la oreja; esta tarea se realiza previo a todo el proceso de reconocimiento, incluso previo a aplicar el algoritmo IRT.

3.3.6. Método de ventanas deslizantes

Motivado por el excelente desempeño de Zhou et al. [129], modificamos su aproximación buscando que el algoritmo fuera invariable a las rotaciones en el plano. En lugar de una ventana de detección rectangular, usamos una combinación de ventanas circulares y rectangulares de las que se almacenaron un conjunto de coordenadas resultante.

El reconocimiento de imágenes mayores que el tamaño de entrada NN se logra mediante el enfoque de ventana deslizante (ver figura 3.3.4). Este algoritmo se define por dos valores, el tamaño de la ventana S, usualmente fijado para coincidir con el tamaño de entrada diseñado en la NN; y el paso entre ventanas T, que especifica la distancia a la que ventanas consecutivas son espaciadas una de otra. Este paso establece el total de ventanas analizadas, W, para una imagen de entrada dada. Para una imagen de tamaño $I_w \times I_h$, la cantidad de ventanas está dada por la formula 3.10.

$$W = \left(\frac{I_w - S}{T} + 1\right)\left(\frac{I_h - S}{T} + 1\right) \quad \Longrightarrow \quad W \propto \frac{I_w I_h}{T^2} \tag{3.10}$$

La figura 3.3.4 muestra la aplicación de este método en una imagen de entrada, extrayendo ventanas de S igual a 32 para este caso simple donde $T = S/2$. Una red analizando esta imagen requeriría 40 ejecuciones para completar el análisis de todas las ventanas extraídas. El requerimiento computacional crece si se selecciona un paso entre ventanas más pequeño. Una acción necesaria para mejorar la velocidad de resolución del clasificador es $T = S/8$.

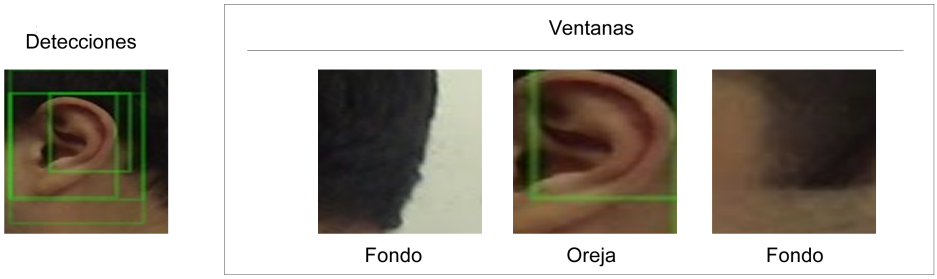


Figura 3.3.4: Vistazo al método de ventanas deslizantes

3.3.7. Configuración experimental

Para probar el resultado de nuestro enfoque, básicamente probamos la red neuronal profunda frente a una red neuronal estándar de alimentación hacia adelante fusionada con algoritmos tradicionales como *principal components analysis (PCA)*, *lineal discriminant analysis (LDA)* y *speeded up robust features (SURF)*. De hecho, utilizamos nuestra colección de vídeos como fuente de entrada (90 segundos y 30 frames x segundo), los cuales fueron obtenidos en ambientes controlados de iluminación y perspectiva. Esta tarea nos proporcionó 118.800 orejas de 11 sujetos distintos.

La configuración de parámetros de las redes neuronales utilizada en esta investigación es dinámica: la salida de las neuronas depende del filtro utilizado por la distancia Hausdorff quien determina cómo se construirá la red neuronal y que usuarios serán finalmente representados en las neuronas de salida; con el fin de reducir la cantidad de posibles respuestas, es decir, de mil neuronas de salida equivalente a mil usuarios en la base de datos, tendremos una reducción a por ejemplo diez neuronas de salida si la distancia Hausdorff nos indica que solo diez orejas superan el umbral de similitud definido en este algoritmo. La capa oculta de la red estándar se construye dinámicamente, respetando que el número de neuronas debería ser entre la cantidad de neuronas de entrada y el tamaño de la capa de salida, es decir, 2/3 del tamaño de la capa de entrada, más la capa de salida; y menos del doble de la capa de entrada, basado en la investigación de Jeff Heaton [99].

3.3.7.1. Análisis de componentes principales

El algoritmo de reconocimiento de orejas con *eigenears* es descrito basicamente diciendo que las imagenes originales del conjunto de entrenamiento son transformadas en un conjunto de eigenears E, entonces, los pesos de cada imagen en el conjunto (E) se calculan, y se almacenan en otro conjunto W. Observando la imagen desconocida X, se le calculan de igual manera sus pesos, y se construye un vector con sus pesos W_X. Subsiguientemente, W_X es comparado con los pesos de las imágenes [69].

El proceso de clasificar una nueva oreja en Γ_{new} a otra clase (orejas conocidas) es el resultado de dos momentos. Primero, la nueva imagen es transformada en sus componentes *eigenear*. Los pesos resultantes forman el vector de pesos Ω_{new}^T.

$$\begin{aligned} \omega_k &= u_k^T(\Gamma_{new} - \Psi) \;\; k = 1, \ldots, M' \\ \Omega_{new}^T &= [\omega_1 \; \omega_2 \; \ldots \; \omega_{M'}] \end{aligned} \tag{3.11}$$

La distancia euclidea entre los dos vectores $d(\Omega_i, \Omega_j)$ proporciona una medida de similitud entre las correspondientes imágenes i y j. Si la distancia entre Γ_{new} y el resto de las imagenes en promedio excede un cierto umbral, se puede asumir que Γ_{new} no es una oreja reconocible [69]. Después los vectores eigen son calculados y almacenados para utilizarlos como entrada de la red estándar de alimentación hacia adelante.

3.3.7.2. Análisis discriminante lineal

LDA o fisherears, supera las limitaciones del algoritmo PCA aplicando el criterio de discriminación lineal de fisher. El método PCA es una combinación lineal de funciones que maximiza la varianza de la información. Esto resulta en un desempeño pobre, especialmente cuando trabajamos con imágenes con demasiado ruido como cambios en el fondo, luz y perspectiva. Entonces el algoritmo PCA puede encontrar componentes defectuosos para la clasificación. Para evitar estos inconvenientes, se implementó el algoritmo de fisher a fin de comparar los resultados. El algoritmo básicamente se implementa siguiendo lo expuesto por Wagner et al [190].

Construimos la matriz de imágenes x con cada columna representando una imagen. Cada imagen se asigna a una clase en su correspondiente vector de clases c. Se proyecta x en el sub-espacio dimensional $(N - c)$ como P con la matriz rotada $WPca$ identificada como un PCA, donde N es el numero de muestras en x, c es un numero único de clases $(length(unique(C)))$ y calculamos la distancia entre clases de la proyección P como se muestra en la formula 3.12.

$$Sb = \sum_{i=1}^{c} N_i * (mean_i - mean) * (mean_i - mean)^T \tag{3.12}$$

Donde $mean$ es el total medio de P, $mean_i$ es la media de la clase i en P, N_i es el número de muestras por clase i. Entonces, necesitamos calcular la distancia entre clases de P, esto se detalla en la ecuación 3.13.

$$Sw = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - mean_i) * (x_k - mean_i)^T \tag{3.13}$$

Donde x_i son las muestras de clase i, x_k es una muestra de x_i, $mean_i$ es la media de la clase i en P. Aplicamos un análisis discriminante estándar lineal y maximizamos el ratio del determinante de dispersión entre clases y dispersión dentro de la clase. La solución es dada por el conjunto generalizado de eigenvectors, *Wfld*, de *Sb* y *Sw* correspondiente a su *eigenvalue*. El rango de *Sb* es como mucho $(c - 1)$, entonces hay solamente $(c - 1)$ no cero eigenvalues, separados del resto. Finalmente obtenemos las fisherears con $W = WPca * Wfld$ [190]. Estos vectores son utilizados como entrada para entrenar la red neuronal de la misma manera como se explicó en la sección precedente.

3.3.7.3. Speeded up robust features (SURF)

La imagen de la oreja es recreada a través del algoritmo SURF como un conjunto de puntos salientes, donde cada uno es asociado con un vector descriptor y que puede ser de una dimensión de 64 o 128. Los vectores descriptores de dimensión 128 son considerados como los que contienen mayor exactitud en las características representadas de una imagen esto es que almacenan características suficientemente descriptivas. El método para identificar de forma única a un individuo se propone combinando las características obtenidas de varias instancias de entrenamiento.

Si tenemos n imágenes de la oreja de un individuo para entrenamiento, un prototipo fusionado se obtiene de combinar el arreglo de descriptores de toda la colección, considerando los vectores redundantes únicamente una vez. Con todas las imágenes procesadas, se construye una colección de etiquetas de a quien pertenece cada imagen y cada vector fusionado, calculados previamente. Después se calculan las características SURF, y se filtran las imágenes a través de la distancia Hausdorff y las características unidireccionales de las imágenes son depositadas en la base de datos.

Estos vectores son utilizados como entrada para entrenar la red estándar. En la etapa de entrenamiento, las matrices unidireccionales de valores pertenecientes a un individuo se toman como positivas retornando el valor uno en la neurona de salida y cero en las otras neuronas. Cuando la nueva imagen ha sido capturada, se recalcula el vector de características con el algoritmo SURF, y con el vector de la imagen marcada como desconocida se procede a intentar reconocerlo a través de la red neuronal. Si la salida supera el umbral definido, entonces se determina que la imagen pertenece al usuario donde se activó la neurona correspondiente.

3.3.8. Resultados experimentales

La configuración de los experimentos en la presente investigación fue realizada con el propósito de determinar si es posible utilizar redes neuronales profundas en el reconocimiento de personas a partir de imágenes de las orejas. Los resultados preliminares son alentadores, las tablas 3.3.1 y 3.3.2 muestran respectivamente la matriz de confusión de entrenamiento y prueba de las 170 redes neuronales convolucionales, construidas para el mismo número de individuos las cuales realizaron reconocimiento entre un enorme conjunto de datos negativos.

En cada matriz, puede observarse una diferencia significativa entre los usuarios clasificados correctamente versus los falsos negativos y falsos positivos. Aunque los datos son alentadores, es importante hacer notar que el entrenamiento de las redes neuronales profundas fue realizado con imágenes bajo condiciones similares de iluminación y perspectiva, y a la vez con la misma calidad, y similar tamaño de orejas; resumiendo, sin ningún tipo de ruido. Esta configuración no determina la efectividad del sistema en ambientes no colaborativos.

Tabla 3.3.1: Matriz de confusión durante entrenamiento

Clasificado / Clase real	Identificado correctamente	Fondo	Total en clase	Exactitud (%)
Identificado correctamente	26071	329	26400	98, 75
Fondo	190	26210	26400	99, 28
		Total	52800	99, 02

Tabla 3.3.2: Matriz de confusión durante etapa de prueba

Clasificado / Clase real	Identificado correctamente	Fondo	Total en clase	Exactitud (%)
Identificado correctamente	12890	310	13200	97, 65
Fondo	211	12989	13200	98, 40
		Total	26400	98, 03

Tabla 3.3.3: Desempeño de nuestra red profunda (CNN) versus PCA, LDA y SURF-NN

	Conjunto de datos Escuela de Policía de Ávila				Conjunto de vídeos de Bisite			
	Precisión (%)	Recuperación (%)	Exactitud (%)	Métrica F1 (%)	Precisión (%)	Recuperación (%)	Exactitud (%)	F1 Score (%)
CNN	97,71	80,85	84,82	88,48	79, 17	43, 65	40, 13	56, 27
PCA-NN	98,00	61,03	66,37	75,22	48, 38	28, 46	21, 83	35, 84
LDA-NN	98,40	69,13	68,36	81,21	52, 31	36, 47	27, 37	42, 98
SURF-NN	98,95	77,39	76,75	86,85	53, 18	42, 03	30, 68	46, 95

Como parte de la evaluación del proceso se tomaron 44 vídeos, que comprimen 90 segundos en promedio, donde aproximadamente 60 segundos mantienen las condiciones descritas anteriormente, y 30 segundos presentan cambios de perspectiva; cada vídeo tiene 30 frames por segundo. Es lógico imaginar que el uso de imágenes bajo las condiciones descritas afecte a todos los algoritmos que se han probado. La tabla 3.3.3 muestra un análisis comparativo entre los dos conjuntos de datos utilizados en la investigación, donde se puede observar claramente cómo la medida F1 decrece radicalmente cuando se utiliza un conjunto aleatorio de las imágenes de los vídeos. Lo anterior nos hace pensar que para la correcta operación de las redes neuronales profundas es necesario entrenarlas con datos que deben ser tan similares como sea posible tomando en consideración los futuros objetos a reconocer.

Es importante resaltar las diferencias entre la medida F1 en condiciones ideales y cuando el conjunto de datos es altamente difuso. El resultado no es del todo significativo. Sin embargo, mantiene una diferencia positiva, cuando las imágenes son cercanas a lo que se espera del mundo real y aunque su precisión puede considerarse decepcionante, muestra que tiene un importante potencial de mejora.

3.3.9. Conclusiones

A lo largo del documento hemos presentado un proceso de reconocimiento y detección utilizando redes neuronales convolucionales. Las CNN's trabajan utilizando la perceción por imagen tienden a ser un enfoque bastante robusto en comparación a los sistemas tradicionales de visión por ordenador que confían en características recortadas manualmente para cada tarea específica. Especialmente cuando se prueba este tipo de sistemas en ambientes donde las variables presentan problemas como la oclusión o el ruido, detalles que son comunes en las imágenes obtenidas del entorno y que representan lo que la oreja de una persona es en el mundo real. Sería importante lograr alcanzar la tasa de efectividad de los algoritmos de detección en el proceso de reconocimiento, lo que hemos observado es que si bien la oreja puede detectarse con bastante precisión, el reconocimiento no termina de ser suficientemente exacto en condiciones no controladas.

La presente investigación considera como propias las modificaciones de algoritmos existentes, la fusión de técnicas utilizadas en el reconocimiento facial y transportadas a la oreja como característica biométrica y la implementación de los enfoques propios, así como el análisis comparativo entre estos algoritmos.

El uso de redes neuronales profundas está cobrando especial relevancia en la identificación mediante elementos biométricos y su uso en el reconocimiento de orejas ha demostrado a ser prometedor. Por tanto las lineas futuras de investigación en el reconocimiento van orientadas a mejorar el desempeño en ambientes no controlados, imágenes en tiempo real y en un entorno del día a día. La utilidad de la presente investigación va más alla de los resultados, demostrando la viabilidad del uso de redes profundas no sólo en el reconocimiento sino también en la detección de objetos. En futuras investigaciones se intentará obtener acceso a nuevos conjuntos de datos públicos, y convertir las múltiples redes en una única red convolutional que combine todos los procesos aquí descritos.

4

Conclusiones y líneas de trabajo futuras

4.1. Resumen de resultados y conclusiones

Hemos experimentado diferentes métodos para detectar la región de la oreja en el perfil del rostro. Además, hemos evaluado el impacto de proyectar la imagen a distintas escalas utilizando un enfoque de ventana deslizante y extrayendo las características de la oreja a partir de distintos algoritmos.

La investigación ha girado en torno a un nuevo enfoque para la detección y reconocimiento de orejas basado en algoritmos de procesamiento de imágenes y redes neuronales convolucionales. Los experimentos han concluido que en condiciones ideales el utilizar 1-CNN en la etapa de detección es altamente eficiente, llegando incluso a obtener una exactitud del $99,85\,\%$, de igual forma, el utilizar 3-CNN's en similares condiciones controladas, muestra una precisión estable superior al $99\,\%$.

Posteriormente evaluamos las redes profundas en conjuntos de datos con imágenes del mundo real, ambiente no controlado con condiciones variables de iluminación, ruido y oclusión, los resultados alcanzados demuestran que el utilizar una sola red neuronal profunda no es suficiente para superar las dificultades que estos conjuntos representan, se alcanzó una exactitud de tan sólo $41,46\,\%$. En condiciones idénticas la integración de 3-CNN's logró un valor de métrica F1 del $85,77\,\%$.

En el proceso de reconocimiento, el destacar las características tubulares de la oreja con el algoritmo IRT facilitó la identificación de los elementos clave que hacen de la oreja un rasgo biométrico único. Posteriormente estas características servirían de entrada a la red neuronal profunda. Los resultados de los experimentos demuestran que en conjuntos de datos no controlados las redes neuronales convolucionales realizan un mejor trabajo en el proceso de reconocimiento incluso superando las mejoras que se realizaron a algoritmos tradicionales como PCA, LDA, y SURF a los que se les integraron redes neuronales de alimentación hacia adelante para mejorar sus resultados. En los siguientes apartados se detallan las conclusiones alcanzadas y los resultados obtenidos en las etapas de detección y reconocimiento.

4.1.1. Detección

Los métodos propuestos para la detección de orejas en vídeos e imágenes 2D son (i) el uso de redes convolucionales (sección 3.2) y (ii) la modificación del algoritmo de rayos para resaltar las propiedades tubulares de la oreja (sección 3.3.3). Hemos llegado a la conclusión que la estructura superficial de la forma de la oreja es única, a tal grado, que un pequeño conjunto de reglas simples es suficiente para detectar qué región la contiene y diferenciarla del resto de la imagen del perfil del rostro. Este conjunto de reglas describe las propiedades que se pueden extraer de un mapa de profundidad observando alta curvatura superficial. El sistema se ha diseñado de tal manera que es invariante a rotaciones en el plano y robusto ante la

presencia de variaciones y oclusión parcial. También mostramos que este enfoque puede refinarse aún más, combinando información de los mapas de bordes y superficie del objeto observado. Al comparar enfoques similares se debe tener especial cautela ya que dichos enfoques pudieron ser evaluados en diferentes conjuntos de datos, con expectativas distintas y bajo diferentes restricciones para una detección exitosa. Nuestro enfoque en detección tiene una tasa de precisión similar a muchos de los presentados en el capítulo 1 con la particularidad que su desempeño bajo condiciones no colaborativas de oclusión, perspectiva e iluminación le otorgan un distintivo adicional necesario en cada investigación.

4.1.2. Segmentación a partir de imágenes de perfil con un enfoque de ventana deslizante

Para detectar la oreja en imágenes de gran tamaño modificamos el método tradicional de ventana deslizante, dónde, En lugar de una ventana de detección rectangular, usamos una combinación de ventanas circulares y rectangulares (ver sección3.3.6). El objetivo era proyectar la imagen 2D en un plano de coordenadas polares 3D a partir del vector de características de la ventana circular y así detectar la oreja independientemente del ángulo de rotación, sin embargo, sólo fue parcialmente alcanzado y a un costo computacional bastante elevado.

Además del alto costo computacional, el rendimiento de detección no fue satisfactorio. El efecto inicial fue una pérdida de datos durante la proyección al espacio cartesiano de coordenadas polares, lo que resulto en una tasa de detección más baja. En un análisis posterior, entendimos que proyectar la oreja en coordenadas polares hacía que los píxeles de las secciones interiores de la oreja fueran sobre-representados, es decir, que la proyección exageraba los detalles internos de la oreja restando importancia a las secciones exteriores. Además los píxeles de los bordes externos se combinaban calculando su representación como el promedio de varios píxeles circundantes. Por lo tanto, perdemos información durante la pro-

yección de los píxeles exteriores y sobre-valoramos las regiones interiores. Como resultado, el vector de características contenía menos información distintiva que el descriptor rectangular y por lo tanto realizaba peor trabajo. Lo que nos llevo a la recomendación de integrar las ventanas circulares con ventanas de detección rectangulares y la aplicación de una normalización, la que se describe en el siguiente apartado, renunciando a la proyección de la imagen 2D en el plano polar.

4.1.3. Normalización de orejas en imágenes recortadas

Como parte de las actividades de procesamiento y a fin de mejorar el rendimiento de los métodos aplicados, procedimos a validar qué algoritmos de preprocesamiento de imágenes podrían facilitar la etapa de reconocimiento, concluyendo que independientemente de que nuestra detección fuere o no invariante ante rotación y perspectiva era requerida una normalización (ver sección 3.3.1).

Los altos costes computacionales nos llevaron a realizar una etapa de filtrado previa a construir la red neuronal profunda que realizaría el reconocimiento final. Esto modificando en diferentes estadios el uso de la distancia euclídea tradicional entre píxeles, por la distancia Hausdorff (ver sección 3.3.4), es decir, en lugar de sólo comparar la posición de un píxel contra otro en dos imágenes, realizamos una comparación de ese píxel en relación al conjunto de píxeles que le rodean y frente al píxel equivalente en la segunda imagen; esto es modificar los algoritmos PCA y LDA reemplazando la distancia euclídea utilizada en sus ecuaciones por Hausdorff y además utilizar este algoritmo como etapa de filtrado definiendo un umbral que descartaba aquellas imágenes demasiado diferentes a la entrada.

4.1.4. Detecciones en condiciones reales

Nuestra investigación sobre la segmentación y detección de la oreja concluye con un estudio sobre la robustez de determinada arquitectura de redes neuronales profundas, combinada con un detector basado en ventanas deslizantes, para señalar la capacidad del detector de identificar una oreja bajo condiciones no cooperativas de oclusión, degradación de imagen, y perspectiva.

Hemos entrenado otros detectores como por ejemplo: Adaboost usando características Haar, redes neuronales tradicionales de alimentación hacia adelante, modificaciones de estas con distancia Hausdorff en lugar de distancia euclídea, etc. Los distintos detectores se aplicaron a copias de la base de datos, donde añadimos ruido y desenfoque a cada imagen.

En la literatura revisada a la fecha no hemos encontrado un estudio que detalle de forma significativa el impacto de la calidad de la imagen y la oclusión en la detección de orejas. Nuestra conclusión es que las redes convolucionales son la mejor opción si la imagen se ha degradado por el ruido y el desenfoque y/o por sólo contar con propiedades parciales producto de la oclusión. Se repitió el mismo experimento, es decir, mismo particionamiento, mismos detectores, mismas métricas de error en una base de datos de 2000 imágenes de orejas en condiciones del mundo real. Para las imágenes degradadas, obtenemos una tasa de error promedio más alta que para las imágenes donde sólo se presenta oclusión. A diferencia de los experimentos con ruido y desenfoque, la detección en escenarios con la mitad de resolución original es a veces ligeramente mejor comparado con el experimento donde se utiliza la resolución completa de las imágenes. De nuevo, las redes profundas demuestran ser una buena opción obteniendo una tasa F1 de detección satisfactoria, aproximadamente del $86,77\,\%$ en condiciones no ideales en comparación al resto de algoritmos cuya métrica F1 oscilaba entre $49,22\,\%$ y $64,48\,\%$.

4.1.5. Reconocimiento

Durante el trabajo de investigación hemos evaluado diferentes enfoques para el reconocimiento, incluyendo descriptores de campo de fuerza, plantillas deformables, aproximaciones geométricas, algoritmos lineales, puntos de interés o descriptores locales, etc. Ninguno de estos métodos fue capaz de darnos estabilidad y repetibilidad, aunque algunos autores publicaron buenas tasas de rendimiento en sus conjuntos de datos. Nuestras aproximaciones con estos algoritmos eran vulnerables a la oclusión y planteaban variaciones significativas en sus resultados en diferentes conjuntos de prueba. Algunos de estos modelos requieren una cantidad suficientemente grande de datos de entrenamiento, que no tenemos. Incluso después de inicializar cuidadosamente una imagen de oreja normalizada, los resultados eran inestables y los bordes extraídos tenían una fuerte tendencia a descartar la estructura tubular de la oreja.

Hemos intentado utilizar clusters de superficie desde el paso de detección para la extracción de las características clave, como el trago y el hélix. Este enfoque funcionó en parte, pero el número de elementos clave fue demasiado pequeño y su posición no es identificable para cada situación, por ejemplo, en ocasiones la calidad de imagen no permite detectar la posición del hélix, por tanto no es lo suficientemente estable.

Después de múltiples intentos fallidos, decidimos omitir la detección por referencia y utilizar las características de textura y superficie en su lugar. En el apartado 3.3, proporcionamos un estudio sobre el rendimiento de los descriptores de características y proponemos un descriptor combinado de texturas que utiliza información 2D y profundidad, esto a partir del algoritmo IRT para detectar de mejor forma la estructura tubular de la oreja. Nuestro descriptor, sin embargo, no pudo superar los enfoques de referencia en imágenes 2D. La investigación confirma que los parámetros de un método pueden ser dirigidos hacia un determinado conjunto de datos con el fin de obtener resultados perfectos en esa base de datos particular;

La misma configuración en otro conjunto de datos, sin embargo, puede producir resultados diferentes. De esta serie de experimentos, aprendimos a interpretar los resultados con precaución, especialmente si se generaron con un sólo conjunto de datos lo que es un indicativo muy limitado para uno u otro enfoque ya que puede interpretarse que los resultados fueron perfectos para ese conjunto en específico y pueden no ser reproducibles en un conjunto distinto.

Por tanto concluimos nuestra investigación mediante un estudio sobre la robustez del reconocimiento en oclusión en distintos conjuntos de datos. El escenario y los parámetros de este estudio son los mismos que para el estudio sobre detección. De acuerdo con nuestros resultados del capítulo 3, el enfoque de extracción de características más fiable como entrada de una red convolucional para reconocimiento es el IRT binarizado.

Al igual que los resultados de la detección, el rendimiento de todos los descriptores continuaba aproximadamente igual si se reducía el tamaño de la imagen de entrada en tan sólo un cuarto de su tamaño original, con una pérdida mínima de reconocimiento y sacrificaba su desempeño bajo la ausencia de iluminación, modificación de escala y variaciones de pose que al final son condiciones del ambiente comunes en el mundo real; a pesar de esto, el uso de redes profundas siempre mostró un mejor desempeño en estas circunstancias no colaborativas.

Finalmente como se mencionó anteriormente para lograr un reconocimiento casi perfecto se tuvo que manipular el conjunto de datos para intentar utilizar imágenes en posiciones casi perfectas donde incluso el ángulo era controlado, y de igual manera la red convolucional demostró un rendimiento satisfactorio superior a otros algoritmos sin llegar a superar métodos de reconocimiento de referencia como puede ser el algoritmo basado en patrones binarios locales (*LBP - Local Binary Patterns*).

4.2. Trabajo futuro

Los resultados de la evaluación de la investigación indican que el reconocimiento de orejas está bastante lejos de ser un problema resuelto, al menos esta afirmación es precisa si se considera su estudio bajo condiciones realistas. Las tasas de rendimiento en los conjuntos de datos académicos aportados por los diferentes autores referenciados en la bibliografía tienden a ser demasiado buenos. La razón principal es, que las imágenes dentro de un determinado conjunto de datos se recogen normalmente bajo las mismas condiciones y con el mismo dispositivo de captura. Los datos demográficos también suelen ser sesgados, porque son recolectados en las universidades, donde, los conjuntos de datos suelen contener imágenes de personas jóvenes.

El rendimiento de los experimentos en cualquier conjunto de datos construido específicamente para una investigación en esta área sólo sirve como prueba de concepto. El progreso de la investigación depende de la disponibilidad de nuevas y más desafiantes bases de datos, por ejemplo, un conjunto de datos podría contener mayores variaciones demográficas, orejas izquierda y derecha del mismo sujeto, modelos 3D, o imágenes adquiridas durante un período de tiempo más largo. Durante la investigación se intento satisfacer muchos de los preceptos antes expuestos sin embargo aún estamos bastante lejos de construir y/o obtener el conjunto de datos ideal para desarrollar el sistema biométrico perfecto. Debemos ser conscientes que, además de los problemas de privacidad y protección de datos, la recopilación de conjuntos de datos biométricos requiere proyectos de larga duración con suficiente personal para ejecutar la recopilación y preparar los datos para su publicación. Los nuevos conjuntos no sólo acelerarían el trabajo en biometría de la oreja, sino también ofrecería la oportunidad de llevar los sistemas existentes a un estadio superior.

Aparte de la necesidad de conjuntos de datos más desafiantes, algunos aspectos interesantes y aún sin resolver de la biometría de la oreja son:

- Detección fiable de puntos de referencia en imágenes bajo variación y/o oclusión. Especialmente la detección y enmascaramiento de áreas ocluidas sería una contribución importante para el reconocimiento.
- Otras técnicas para acelerar las operaciones de búsqueda en bases de datos. En particular, para el análisis de imágenes forenses y otogramas.
- Evaluar la idoneidad del uso y almacenamiento de descriptores binarios y/o plantillas promedio que permitan realizar búsquedas rápidas.
- Estudiar las diferencias entre la oreja izquierda y derecha del mismo sujeto.
- Una base de datos que nos permita investigar los efectos de envejecimiento en la biometría de la oreja.
- Las diferencias en las orejas de gemelos.
- La coincidencia de imágenes de orejas capturadas en ambiente real contra su otograma equivalente.

El sistema de cámaras Kinect ha llamado la atención de la comunidad científica orientada a visión por ordenador. Impulsado por las nuevas posibilidades de estos dispositivos, la comunidad desarrolló sistemas para el reconocimiento facial utilizando el flujo de vídeo 2D y 3D de la cámara. En el contexto de este trabajo, sería interesante portar ese flujo al reconocimiento de orejas si los futuros dispositivos cuentan con una resolución de profundidad suficiente para capturar con precisión los detalles de la estructura de la oreja.

4.2.1. Nuevas aplicaciones

El siguiente paso hacia un sistema de reconocimiento completamente funcional es explorar las limitaciones de los enfoques en la literatura en diferentes escenarios. Esta investigación no hace más que empezar comparando la tasa de rendimiento que se podría lograr utilizando unas determinadas bases de datos versus otros conjuntos de datos mucho más complejos. Un posible caso de uso para la biometría de la oreja podría ser la autenticación en teléfonos móviles. Podemos pensar en dos escenarios diferentes: (i) explorar la posibilidad de desbloquear automáticamente el teléfono al contestar una llamada telefónica, y (ii) usar impresiones de la oreja capturadas de la pantalla táctil para mantener un nivel mínimo de confianza durante la llamada. En el segundo escenario la pantalla sensible al tacto devolvería un patrón de puntos que dependería de la estructura de la oreja del propietario; este patrón se puede utilizar para autenticar un sujeto mientras realiza una llamada telefónica. Los móviles podrían evaluar periódicamente las impresiones de la oreja para enriquecer su algoritmo de reconocimiento.

Amazon [34] ha recibido una patente estadounidense para esta tecnología de reconocimiento de orejas en móviles, su algoritmo permite escanear la oreja del usuario con la cámara frontal de un teléfono inteligente para desbloquear el dispositivo si se acerca el móvil a la oreja. En su patente Amazon argumenta, que su investigación ha comprobado sin lugar a dudas que la forma de las orejas de una persona es tan única como las huellas dactilares, lo que permitiría a la tecnología establecer con precisión que el usuario que contesta la llamada es efectivamente el dueño del smartphone.

Además, exponen que este mecanismo puede llegar a ser útil en situaciones en las que sólo el propietario del teléfono inteligente tiene que ser capaz de responder a las llamadas entrantes. He incluso asegurar a la persona que inicio la llamada que efectivamente el receptor es a quien se quiere contactar. Adicionalmente, según la patente la tecnología permitiría al móvil ajustar automáticamente el volumen del

altavoz y/o auricular de acuerdo a la distancia entre el teléfono y la oreja del individuo; sin embargo, es importante resaltar que al día de hoy no se aprecia ninguna aplicación comercial de lo expuesto en la patente, a pesar que esta fue introducida en la oficina de patentes americana en el año 2015.

Otra aplicación interesante podría ser el uso de cámaras de alta resolución, la oreja podría utilizarse como control en fronteras, complemento a los sistemas de reconocimiento facial existentes y como información adicional a los documentos de pasaporte, siempre que se incluyera en ellos la información adicional de las orejas. Es valido afirmar que a pesar de las numerosas posibilidades de aplicación en materia de seguridad, aplicaciones de vigilancia, medicina forense, investigaciones criminales o control de fronteras, la investigación existente en el reconocimiento de orejas rara vez ha ido más allá de los ambientes de laboratorio. Esto puede atribuirse principalmente a la enorme variabilidad de la apariencia de la oreja en imágenes capturadas en entornos restringidos. Con los recientes avances en la visión artificial, aprendizaje automático e inteligencia artificial por ejemplo, con el aprendizaje profundo, muchos de los problemas de reconocimiento son en la actualidad resolubles, al menos en cierta medida, incluso en entornos restringidos, lo que antes era demasiado difícil de alcanzar en situaciones reales cada día se está convirtiendo en una fuente viable de datos para el reconocimiento de personas.

Los organizadores del *Unconstrained Ear Recognition Challenge (UERC)* [74] propuesto en el congreso IJCB - Octubre 2017 - *(International Joint Conference on Biometrics)* quieren construir sobre los avances descritos anteriormente y abordar el problema del reconocimiento de la oreja "en la naturaleza". El objetivo del reto es avanzar en el estado de la tecnología en el campo del reconocimiento automático, para proporcionar a los participantes un problema de investigación desafiante e introducir un conjunto de datos como punto de referencia para evaluar las últimas técnicas, modelos y algoritmos relacionados con el reconocimiento de orejas en condiciones naturales, por tanto observamos que la investigación en este campo está más activa que nunca.

4.3. Consideraciones finales

El reconocimiento de orejas es una característica biométrica prometedora. Nuestra investigación demuestra que podemos lograr un alto rendimiento de reconocimiento en condiciones ideales superior al 90 % y en la "naturaleza"superior al 80 %. Estos sistemas pueden integrarse fácilmente con el reconocimiento facial existente sin la necesidad de que el usuario interactúe. La singularidad de la oreja, así como la diferencia entre orejas izquierda y derecha, representa información valiosa que se debe utilizar para enriquecer el rendimiento de otros sistemas, especialmente en escenarios donde se cuenta con datos del perfil del rostro.

En la presente investigación se proporcionó una descripción detallada del trabajo existente y se abordaron varios desafíos abiertos en la materia. Investigamos el valor de la estructura tubular de la oreja durante la segmentación, detección y reconocimiento, para finalmente proponer dos algoritmos y combinar la profundidad y la textura tanto para la detección como el reconocimiento. También hemos demostrado que la normalización de las imágenes favorece en gran medida el proceso de identificación, y que un buen algoritmo de detección bajo condiciones no colaborativas es posible.

Además, proporcionamos resultados sobre la robustez de diferentes segmentaciones y técnicas de reconocimiento. Proporcionamos la posibilidad de recuperación rápida de la identidad en conjuntos de datos grandes utilizando etapas de filtrado, lo cual es crucial para cualquier característica biométrica y no ha sido previamente abordado en su totalidad para el reconocimiento de orejas. Para realizar una búsqueda secuencial se pueden utilizar representaciones binarias de los descriptores y utilizarlos para construir una nueva proyeccción del objeto en bases de datos de gran tamaño. Se demuestra que la tasa de identificación verdadero-positiva de un sistema que utiliza el pre-procesamiento y filtrado es superior a aquellos que no realizan esta tarea. Concluimos que el reconocimiento con redes profundas es posible, completamente viable y altamente efectivo.

4.4. Listado de publicaciones

Finalmente, presentamos las publicaciones que a la fecha se han derivado de la investigación.

1. Ear detection and localization with convolutional neural networks in natural images and videos[1] *[Journal of Machine Vision and Applications, 2016]*.
2. A brief review of the ear recognition process using deep neural networks *[Journal of Applied Logic, JAL 2016]*. Factor de impacto: 0.524
3. Exploring the Ear Recognition Process *[International Journal of Imaging and Robotics, IJIR 2015]*
4. A Small Look at the Ear Recognition Process using a Hybrid Approach *[Journal of Applied Logic, JAL 2015]*. Factor de impacto: 0.524
5. Ear Recognition using a Hybrid Approach based on Neural Networks *[17th International Conference on Information Fusion, FUSION'14]*
6. Neural Networks using Hausdorff Distance, SURF and Fisher Algorithms for Ear Recognition *[9th International Conference Soft Computing Models in Industrial and Environmental Applications, SOCO'14]*
7. Ear Recognition with Neural Networks based on Fisher and Surf algorithms *[International Conference on Hybrid Artificial Intelligence Systems, HAIS'14]*
8. A Brief Approach to the Ear Recognition Process *[Advances in Intelligent Systems and Computing, 2014, DCAI'14]*
9. Ear Biometrics: A Small Look at the Process of Ear Recognition *[Soft Computing Models in Industrial and Environmental Applications, 2013, SOCO'13]*
10. Face Identification by Real-Time Connectionist System, *[Advances in Intelligent Systems and Computing Volume 217, 2013 pp 393-400, DCAI'13]*

[1] En revisión

Bibliografía y referencias

[1] **ISO/IEC 2382-37**. General biometric system. *International Organization for Standardization*, 2012. Citado en las páginas 12 y 167.

[2] **Abate A., Nappi M., Riccioand D., and Ricciardi S.** Ear recognition by means of a rotation invariant descriptor. *18th International Conference on Pattern Recognition (ICPR)*, 2006. Citado en página 39.

[3] **Abaza A. and Ross A.** Towards understanding the symmetry of human ears: A biometric perspective. *IEEE International Conference on Biometrics: Theory Applications and Systems*, 2010. Citado en página 19.

[4] **Abaza A. and Harrison M. A. F.** Ear recognition: a complete system. *Biometric and Surveillance Technology for Human and Activity Identification*, 2013.

[5] **Abaza A., Hebert C., and Harrison M.** Fast learning ear detection for real-time surveillance. *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems*, 2010. Citado en las páginas 30, 33, y 64.

[6] **Abaza A., Ross A., Hebert C., Harrison M. A. F., and Nixon M. S.** A survey on ear biometrics. *ACM Computing Surveys*, 2013. Citado en las páginas 7, 21, 32, y 35.

[7] **Al Nizami H. A., Adkins-Hill J. P., Zhang Y., J. R. Sullins, C. McCullough, and S. Canavanand L. Yin**. Fast learning ear detection for real-time surveillance. *A biometric database with rotating head videos and hand-drawn face sketches*, 2009. Citado en página 27.

[8] **Anjos A. and Marcel S.** Counter-measures to photo attacks in face recognition: a public database and a baseline. *International Joint Conference on Biometrics (IJCB)*, 2011.

[9] **Bertillon A.** La photographie judiciaire: Avec un appendice sur la classification. *Et La'Identification Anthropometriques*, 1890. Citado en las páginas 7 y 19.

[10] **Conan Doyle A.** A chapter on ears. *Solis Press*, 2012. Citado en página 7.

[11] **Cummings A., Nixon M., and Carter J.** A novel ray analogy for enrolment of ear biometrics. *In Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, 2010. Citado en las páginas 30, 36, y 105.

[12] **Hoogstrate A., Heuvel H. V. D., and Huyben E.** Ear identification based on surveillance camera images. *Science & Justice*, 2001. Citado en las páginas 6 y 20.

[13] **Jain A., Ross A., and Prabhakar S.** An introduction to biometric recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2004.

[14] **Kong A., Zhang D., and Kamel M.** A survey of palmprint recognition. *Pattern Recognition*, 2009.

[15] **Kumar A. and Wu C.** Automated human identification using ear imaging. *International Conference on Pattern Recognition*, 2012. Citado en las páginas 25, 38, 40, y 57.

[16] **Kumar A. and Chan T. S. T.** Robust ear identification using sparse representation of local texture descriptors. *Pattern Recognition*, 2012.

[17] **Kumar A., Hanmandlu M., Kuldeep M., and Gupta H.** Automatic ear detection for online biometric applications. *In Third National Conference on Computer Vision*, 2011. Citado en las páginas 30, 36, y 40.

[18] **Mhatre A., Palla S., Chikkerur S., and Govindaraju V.** Efficient search and retrieval in biometric databases. *In SPIE Defense and Securiry Symposium*, 2005. Citado en página 7.

[19] **Ogale N. A.** A survey of techniques for human detection from video. *Survey, University of Maryland*, 2006.

[20] **Pflug A. and Busch C.** Ear biometrics: a survey of detection, feature extraction and recognition methods. *Biometrics, IET*, 2012. Citado en página 21.

[21] **Pflug A. and Busch C.** Segmentation and normalization of human ears using cascaded pose regression. *In 19th Nordic Conference on Secure IT Systems*, 2014. Citado en las páginas 13 y 14.

[22] **Pflug A., Winterstein A., and Busch C.** Ear detection in 3d profile images based on surface curvature. *In Proceedings of IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2012. Citado en página 14.

[23] **Pflug A., Hartung D., and Busch C.** Feature extraction from vein images using spatial information and chain codes. *Human Factors and Biometrics.*, 2012.

[24] **Pflug A., Back P., and Busch C.** Towards an ear detection that is robust against rotation. *In The 46th Annual IEEE International Carnahan Conference on Security Technology*, 2012. Citado en página 13.

[25] **Pflug A., Winterstein A., and Busch C.** Robust localization of ears by feature level fusion and context information. *In Proceedings of the International Conference on Bio-metrics (ICB)*, 2013.

[26] **Pflug A., Ross A., and Busch C.** 2d ear classification based on unsupervised clustering. *In Proceedings of International Joint Conference on Biometrics (IJCB)*, 2014.

[27] **Pflug A., Wagner J., Rathgeb C., and Busch C.** Impact of severe signal degradation on ear recognition performance. *In Proceedings of Biometrics & Forensics & De-identification and Privacy Protection (BiForD)*, 2014.

[28] **Pflug A., Paul P., and Busch C.** A comparative study on texture and surface descriptors for ear biometrics. *In Proceedings of the International Carnahan Conference on Security Technology (ICCST)*, 2014.

[29] **Pflug A., Rathgeb C., and Scherhag U.** Binarization of histogram models: An application to efficient biometric identification. *In Proceedings of International Conference on Cybernetics*, 2015.

[30] **Ross A. and Mukherjee R.** Augmenting ridge curves with minutiae triplets for fingerprint indexing. *In SPIE Biometrics*, 2007.

[31] **Sana A., Gupta P., and Purkait R.** Ear biometrics: A new approach. *In Advances in Pattern Recognition*, 2007. Citado en las páginas 39 y 42.

[32] **Scheenstra A., Ruifrok A., and Veltkamp R.** A survey of 3d face recognition methods. *In Audio and Video Based Biometric Person Authentication*, 2005.

[33] **Yazdanpanah A. and Faez K.** Normalizing human ear in proportion to size and rotation. *In Emerging Intelligent Computing Technology and Applications of Lecture Notes in Computer Science*, 2009.

[34] **Amazon**. Ear recognition technology for smartphones, 2015. URL https://www.uspto.gov/. Citado en página 130.

[35] **Asmaa Sabet Anwar, Kareem Kamal A. Ghany, and Hesham Elmahdy**. Human ear recognition using sift features. *Third World Conference on Complex Systems (WCCS)*, 2015.

[36] **Asmaa Sabet Anwar, Kareem Kamal A. Ghany, and Hesham Elmahdy**. Human ear recognition using geometrical features extraction. *Procedia Computer Science*, 2015.

[37] **Arbab-Zavar B. and Nixon M.** On shape-mediated enrolment in ear biometrics. *Advances in Visual Computing*, 2007. Citado en las páginas 30, 35, y 43.

[38] **Arbab-Zavar B. and Nixon M.** Robust log-gabor filter for ear biometrics. *In International Conference on Pattern Recognition (ICPR)*, 2008. Citado en las páginas 30, 39, y 41.

[39] **Arbab-Zavar B., Nixon M., and Hurley D.** On model-based analysis of ear biometrics. *In First IEEE International Conference on Biometrics: Theory, Applicationsand Systems*, 2007. Citado en las páginas 39 y 44.

[40] **Bhanu B. and Chen H**. Human ear recognition by computer. *3D Ear Detection from Side Face Range Images*, 2008.

[41] **Moreno B., Sanchez A., and Velez J.** On the use of outer ear images for personal identification in security applications. *In IEEE 33rd Annual 1999 International Carnahan Conference on Security Technology*, 2002. Citado en página 39.

[42] **Scholkopf B., Smola A., and Mueller K. R.** Kernel principal component analysis. *In Advances in Kernel Methods - Support Vector Learning*, 1999.

[43] **Victor B., Bowyer K., and Sarkar S.** An evaluation of face and ear biometrics. *In 16th International Conference on Pattern Recognition (ICPR)*, 2002. Citado en las páginas 39 y 47.

[44] **Zhang B., Mu H., and Zeng Z.** Ear recognition based on gabor scale information. *In International Conference on Wavelet Analysis and Pattern Recognition*, 2013. Citado en página 57.

[45] **Zhang B., Mu Z., Zeng H., and Luo S.** Robust ear recognition via nonnegative sparse representation of gabor orientation information. *The Scientific World Journal*, 2014. Citado en las páginas 48, 57, y 59.

[46] **ISO/IEC 19795-1 JTC1 SC37 Biometrics**. Information technology biometric performance testing and reporting part 1: Principles and framework. *International Organization for Standardization and International Electrotechnical Committee*, 2006.

[47] **ISO/IEC JTC1 SC37 Biometrics**. Iso/iec jtc 1/sc 37 n 3663 working draft: Harmonized biometric vocabulary. *International Organization for Standardization*, 2010. Citado en las páginas 168 y 169.

[48] **ISO/IEC TC JTC1 SC37 Biometrics**. Iso/iec 19795-1 information technology biometric performance testing and reporting. *International Organization for Standardization and International Electrotechnical Committee*, 2006.

[49] **G. Bradski**. Opencv. *Dr. Dobb's Journal of Software Tools*, 2000. Citado en las páginas 86 y 91.

[50] **Bundeskriminalamt**. Qualitätstatndardbeschreibung zur fertigung und anzeige von didigital erkennungsdienstlichen llichtbilder. *INPOL*, 2014.

[51] **Busch C. and Brauckmann M.** Towards a more secure border control with 3d face recognition. *In Norsk informasjonssikkerhetskonferanse (NISK)*, 2012.

[52] **Busch C., Pflug A., Zhou X., Dose M., Brauckmann M., Helbig J., Opel A., Neuge bauer P., Leowski K., Sieber H., and Lotz O.** Multibiometrische gesichtserkennung. *13 Deutscher IT-Sicherheitskongress*, 2013.

[53] **Faltemier T. C., Bowyer K. W., and Flynn P. J.** Rotated profile signatures for robust 3d feature detection. *In Automatic Face and Gesture Recognition*, 2008. Citado en página 24.

[54] **Huang C., Camps O., and Kannungo T.** Object recognition using appearance based parts and relations. *In IEEE Conference on Computer Vision and Pattern Recognition*, 1997. Citado en página 3.

[55] **Huang C., Camps O., and Kannungo T.** Hierarchical organization of appearance based parts and relations for object recognition. *In IEEE Conference on Computer Vision and Pattern Recognition*, 1998. Citado en página 3.

[56] **Sanderson C., Bgdeli A., Shan T., Chen S., Berglund E., and Lovell B. C.** Intelligent cctv for mass transport security: Challenges and opportunities for video and face processing. *Electronic Letters on Computer Vision and Image Analysis*, 2007.

[57] **Sforza C., Grandi G., Binelli M., Tommasi D. G., Rosati R., and Ferrario V. F.** Age and sex related changes in the normal human ear. *Forensic Science International*, 2009. Citado en página 20.

[58] **Wu C.** Sift gpu: A gpu implementation of scale invariant feature transform (sift). *http://cs.unc.edu/~ccwu/siftgpu*, 2007.

[59] **Wu C.** Visualsfm: A visual structure from motion system. *http://ccwu.me/vsfm/*, 2011.

[60] **Christopher M. Cyr and Benjamin B. Kimia**. 3d object recognition using shape similarity-base aspect graph. *In IEEE International Conference on Computer Vision*, 2001. Citado en página 6.

[61] **Dimov D. and Cantoni V.** Appearance based 3d object approach to human ears recognition. *Biometric Authen-tication, Lecture Notes in Computer Science*, 2014. Citado en página 59.

[62] **Frejlichowski D. and Tyszkiewicz N.** The west pomeranian university of technology ear database a tool for testing biometric algorithms. *Image Analysis and Recognition*, 2010. Citado en las páginas 24 y 25.

[63] **Hartung D., Pflug A, and Busch C.** Vein pattern recognition using chain codes, spacial information and skeleton fusing. *In GI-Sicherheit*, 2012.

[64] **Kisku D., Mehrotra H., Gupta P., and Sing J.** Sift based ear recognition by fusion of detected keypoints from color similarity slice regions. *In International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, 2009. Citado en las páginas 40 y 44.

[65] **Lowe G. D.** Object recognition from local scale invariant features. *In IEEE International Conference on Computer Vision (ICCV 1999)*, 1999. Citado en página 43.

[66] **Maltoni D., Jain A. Maio D., and S. Prabhakar.** Handbook of fingerprint recognition. *Springer-Verlag, 1st edition*, 2009.

[67] **Zhang D., Guo Z., Lu G., Zhang D., and Zuo W.** An online system of multispectral palmprint verification. *Transaction on Instrumentation and Measurement*, 2010.

[68] **Ciresan D.C., Meier U., Masci J., and Schmidhuber J.** Multi-column deep neural network for traffic sign classification. *Neural Networks*, 2012. Citado en página 75.

[69] **Pissarenko Dimitri**. Eigenface-based facial recognition. *April*, 2002. Citado en página 113.

[70] **Bernstein D. E. and Jackson J. D.** The daubert trilogy in the states. *Law and Economics Working Paper Series*, 2004.

[71] **Gonzalez E., Alvarez L., and Mazorra L.** Normalization and feature extraction on ear images. *In International Carnahan Conference on Security Technology*, 2012.

[72] **Jeges E. and Mt L.** Model based human ear localization and feature extraction. *ICMED*, 2007. Citado en las páginas 30, 39, 45, y 46.

[73] **Levina E. and Bickel P. J.** Maximum likelihood estimation of intrinsic dimension. *In Neural Information Processing Systems*, 2004.

[74] **Ziga Emersic, Vitomir Struc, and Peter Peer**. Unconstrained ear recognition challenge. *International Joint Conference on Biometrics (IJCB)*, 2017. Citado en página 131.

[75] **Abate A. F., Nappi M., Riccioand D., and Sabatino G.** 2d and 3d face recognition: A survey. pattern recognition letters. *International Conference on Pattern Recognition (ICPR)*, 2007. Citado en página 41.

[76] **Cootes T. F., Wheeler G. V., Walker K. N., and Taylor C. J.** Coupled-view active appearance models. *In Proceedings of the British machine vision conference*, 2000.

[77] **Hao F., Daugman J., and Zielinski P.** A fast search algorithm for a large fuzzy database. *IEEE Trans. Information Forensics and Security*, 2008.

[78] **Kiely T. F.** Forensic evidence: Science and the criminal law, second edition. *Ear Impressions*, 2005.

[79] **PPerronnin F. and Dugelay J.-L.** Clustering face images with application to image retrieval in large databases. *In Proc. SPIE Conf. Biometric Technology for Human Identification II*, 2005.

[80] **Badrinath G. and Gupta P.** Feature level fused ear biometric system. *Seventh International Conference on Advances in Pattern Recognition (ICAPR)*, 2009. Citado en las páginas 40 y 43.

[81] **Oxley G.** Forensic human identification. *Recognition and Imagery Analysis*, 2007. Citado en página 9.

[82] **Passalis G., Kakadiaris I., Theoharis T., Toderici G., and Papaioannou T.** Towards fast 3d ear recognition for real-life biometric applications. *In IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2007. Citado en las páginas 52 y 53.

[83] **Robert R. Goldberg and David G. Lowe**. Verification of 3d parametric models in 2d image data. *IEEE Workshop on Computer Vision*, 1987. Citado en página 5.

[84] **Esther Gonzalez, Luis Alvarez, and Luis Mazorra**. Ami: Ear database. *http://www.ctim.es/research_works/ami_ear_database/*, 2015. Citado en las páginas 68 y 101.

[85] **Bay H., Tuytelaars T, and V. Gool L.** Surf: Speeded up robust features. *In Proceedings of the 9th European Conference on Computer Vision*, 2006. Citado en página 45.

[86] **Chen H. and Bhanu B.** Contour matching for 3d ear recognition. *In Proceedings of the Seventh IEEE Workshop on Applications of Computer Vision*, 2005. Citado en las páginas 31, 32, 51, y 52.

[87] **Chen H. and Bhanu B.** Shape model-based 3d ear detection from side face range images. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (CVPR)*, 2005. Citado en las páginas 30 y 31.

[88] **Chen H. and Bhanu B.** Human ear recognition in 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007. Citado en las páginas 31, 32, 51, y 52.

[89] **Freeman H.** On the encoding of arbitrary geometric configurations. *IRE Transactions on Electronic Computer*, 1961.

[90] **Liu H. and Liu D.** Improving adaboost ear detection with skin-color model and multi-template matching. *In 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, 2010. Citado en página 30.

[91] **Liu H. and Zhang D.** Fast 3d point cloud ear identification by slice curve matching. *In 3rd International Conference on Computer Research and Development (ICCRD)*, 2011. Citado en las páginas 52 y 54.

[92] **Liu H. and Yan J.** Multi-view ear shape feature extraction and reconstruction. *In Third International IEEE Conference on Signal Image Technologies and Internet Based System (SITIS)*, 2007. Citado en las páginas 39, 46, y 53.

[93] **Sieber H.** Das gesichtserkennungssystems (ges). *im Bundeskriminalamt. Presentation*, 2011.

[94] **Skibbe H. and Reisert M.** Circular fourier-hog features for rotation invariant object detection in biomedical images. *In IEEE International Symposium on Biomedical Imaging*, 2012.

[95] **Yang H. and Patras I.** Sieving regression forest votes for facial feature detection in the wild. *In Computer Vision (ICCV), IEEE International Conference on*, 2013.

[96] **Zeng H., Dong J.-Y., Mu Z.-C, and Guo Y.** Ear recognition based on 3d keypoint matching. *In IEEE 10th International Conference on Signal Processing (ICSP)*, 2010. Citado en las páginas 52 y 55.

[97] **Zhang H. and Mu Z.** Compound structure classifier system for ear recognition. *In IEEE International Conference on Automation and Logistics*, 2008. Citado en las páginas 40 y 48.

[98] **Shih H.C., Ho C., Chang H., and Wu C. S.** Ear detection based on arc-masking extraction and adaboost polling verification. *In Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, 2009. Citado en las páginas 20 y 30.

[99] **J. Heaton**. Introduction to neural networks for c#, 2nd edition. *Heaton Press*, 2010. Citado en página 112.

[100] **Alberink I. and Ruifrok A**. Performance of the fearid earprint identification system. *Forensic Science International*, 2007. Citado en página 21.

[101] **Biederman I.** Recognition by components: a theory of human understanding. *Psychol. Review*, 1987. Citado en página 3.

[102] **Biederman I.** Recognizing depth-rotated objects: A review of recent research and theory. *Spatial Vision*, 2000. Citado en página 4.

[103] **Biederman I. and Gerhardstein P.** Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 1993. Citado en página 3.

[104] **Kemelmacher-Shlizerman I. and Basri R.** 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2011.

[105] **Matthews I. and Baker S.** Active appearance models revisited. *International Journal of Computer Vision*, 2004.

[106] **Naseem I., Togneri R., and Bennamoun M.** Best practice recommendation for the capture of mugshots. *National Institute for Standardization (NIST)*, 1997.

[107] **Naseem I., Togneri R., and Bennamoun M.** Sparse representation for ear biometrics. *Advances in Visual Computing*, 2008. Citado en las páginas 40 y 50.

[108] **Beis J. and Lowe D.** Shape indexing using approximate nearest-neighbor search in highdimensional spaces. *IEEE Conference on Computer Vision and Pattern Recognition*, 1997. Citado en página 5.

[109] **Besl P. J.** Surfaces in range image understanding. *Springer*, 1988.

[110] **Black M. J. and Jepson A. D.** Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 1998.

[111] **Bustard J. and Nixon M.** Toward unconstrained ear recognition from two-dimensional images. *Systems, Man and Cybernetics, Part A: Systems and Humans*, 2010. Citado en las páginas 40, 43, y 44.

[112] **Canny J.** A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1986.

[113] **Daugman J.** How iris recognition works. *Circuits and Systems for Video Technology,IEEE Transactions on*, 2004.

[114] **Dong J. and Mu. Z.** Multi-pose ear recognition based on force field transformation. *In Second International Symposium on Intelligent Information Technology Application (IITA)*, 2008. Citado en las páginas 38 y 40.

[115] **Flynn P. J., Bowyer K. W., and Phillips P. J.** Assessment of time dependency in face recognition: An initial study. *Audio and Video-Based Biometric Person Authentication*, 2003. Citado en las páginas 68, 69, y 101.

[116] **Gentile J., Ratha N., and Connell J.** An efficient, two-stage iris recognition system. *In Biometrics: Theory, Applicationsand Systems*, 2009.

[117] **Hurley D. J., Nixon M. S., and Carter J. N.** Force field energy functionals for image feature extraction. *Image and Vision Computing*, 2002. Citado en las páginas 37, 39, y 46.

[118] **Kannala J. and Rahtu E.** Bsif: Binarized statistical image features. *In IEEE International Conference on Pattern Recognition*, 2012.

[119] **Klontz J. and Jain A.** A case study on unconstrained facial recognition using the boston marathon bombings suspects. *Computer*, 2013. Citado en página 6.

[120] **Lei J., Zhou J., and Abdel-Mottaleb M.** A novel shape-based interest point descriptor (sip) for 3d ear recognition. *In International Conference on Image Processing (ICIP)*, 2013. Citado en página 59.

[121] **Lei J., Zhou J., Abdel-Mottaleb M., and You X.** Detection, localization and pose classification of ear in 3d face profile images. *In International Conference on Image Processing (ICIP)*, 2013.

[122] **Rouseeuw P. J.** Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journa of Computational and Applied Mathematics*, 1987.

[123] **Wagner J.** Master's thesis. *Hochschule Darmstadt*, 2014.

[124] **Wagner J., Pflug A., Rathgeb C., and Busch C.** Effects of severe signal degradation on ear detection. *In Proc. of 2nd International Workshop on Biometrics and Forensics (IWBF)*, 2014.

[125] **Wright J., Yang A. Y., Ganesh A., Sastry S. S., and Ma Y.** Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2009.

[126] **Xiao J., Baker S., Matthews I., and Kanade T.** Real-time combined 2d+ 3d active appearance models. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[127] **Zhou J., Cadavid S., and Abdel-Mottaleb M.** Histograms of categorized shapes for 3d ear detection. *In International Conference on Biometrics: Theory Applications and Systems*, 2010. Citado en las páginas 13, 31, y 35.

[128] **Zhou J., Cadavid S., and Abdel-Mottaleb M.** A computationally efficient approach to 3d ear recognition employing local and holistic features. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2011. Citado en las páginas 52, 53, y 59.

[129] **Zhou J., Cadavid S., and Abdel-Mottaleb M.** Exploiting color sift features for 2d ear recognition. *In 18th IEEE International Conference on Image Processing (ICIP)*, 2011. Citado en página 111.

[130] **Zhou J., Cadavid S., and Abdel-Mottaleb M.** An efficient 3d ear recognition system employing local and holistic features. *Information Forensics and Security, IEEE Transactions on*, 2012. Citado en página 59.

[131] **Chang K., Bowyer K. W., Sarkar S., and Victor B.** Comparison and combination of ear and face images in appearance based biometrics. *IEEE Transactions in Pattern Anallysis and Machine Intelligence*, 2003. Citado en las páginas 39 y 47.

[132] **Chang K., Bowyer K. W., Sarkar S., and Victor B.** Comparison and combination of ear and face images in appearance based biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003. Citado en las páginas 68, 69, y 101.

[133] **Jain A. K., Prabhakar S., and Hong L.** A multichannel approach to fingerprint classification. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 1999.

[134] **Lammi H. K.** Ear biometics. technical report. *Lappeenranta University of Technology, Department of Information Technology*, 2004. Citado en página 21.

[135] **Pun K. and Moon Y.** Recent advances in ear biometrics. *n Automatic Face and Gesture Recognition, Proceedings. Sixth IEEE International Conference on*, 2004. Citado en página 21.

[136] **Ramesh K. and Rao K.** Pattern extraction methods for ear biometrics - a survey. *In World Congress on Nature Biologically Inspired Computing (NaBIC)*, 2009. Citado en página 21.

[137] **Wang K. and Zhi-Chun**. A 3d ear recognition method based on auricle structural feature. *International Journal of Computer Science Issues (IJCSI)*, 2013. Citado en página 59.

[138] **Zuiderveld K.** Contrast limited adaptive histogram equalization. *Graphics Gems IV*, 1994.

[139] **M. Khamiss, S. Algabary, Khairuddin Omar, Md. Jan Nordin, and Siti Norul Huda S. Abdullah**. Ear identification based on improved algorithm. *17th International Conference on Primate Socioecology and Comparative Methods*, 2015.

[140] **Alvarez L., Gonzalez E., and Mazorra L.** Fitting ear contour using an ovoid model. *39th Annual International Carnahan Conference on Security Technology*, 2005. Citado en las páginas 30 y 34.

[141] **Chen L, Mu Z, Zhang B, and Zhang Y**. Ear recognition from one sample per person. *PLoS ONE*, 2015.

[142] **Gutierrez L., Melin P., and Lopez M.** Modular neural network integrator for human recognition from ear images. *In The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010. Citado en las páginas 40 y 50.

[143] **Kwan-Ho L., Kin-Man L., and Wan-Chi S.** Spatially eigen-weighted hausdorff distance for face recognition. *Hong Kong*, 2002. Citado en página 108.

[144] **Meijerman L.** Inter and intra-individual variation in earprints. *PhD thesis, University of Leiden*, 2006.

[145] **Meijerman L., Sholl S., Conti F. D., Giacon M., Van der Lugt C., Drusini A., Vanezis P., and Maat G.** Exploratory study on classification and individualisation of earprints. *Forensic Science International*, 2004. Citado en página 20.

[146] **Meijerman L., Van Der Lugt C., and Maat G. J.** Cross sectional anthropometric study of the external ear. *Journal of Forensic Sciences*, 2007. Citado en las páginas 7 y 20.

[147] **Nanni L. and Lumini A.** A multi matcher for ear authentication. *Pattern Recognition Letters*, 2007. Citado en las páginas 39 y 49.

[148] **Spreeuwers L.** Fast and accurate 3d face recognition. *International Journal of Computer Vision*, 2011.

[149] **Yuan L. and Mu Z.** Ear recognition based on 2d images. *In First IEEE International Conference on Biometrics: Theory, Applicationsand Systems (BTAS)*, 2007. Citado en las páginas 39 y 47.

[150] **Yuan L. and Mu Z.** Ear recognition based on gabor features and kfda. *The Scientific World Journal*, 2014. Citado en página 57.

[151] **Yuan L. and Mu Z.-C.** Ear detection based on skin-color and contour information. *In International Conference on Machine Learning and Cybernetics*, 2007. Citado en las páginas 30 y 34.

[152] **Yuan L., Chun Mu Z., Zhang Y., and Liu K.** Ear recognition using improved non-negative matrix factorization. *In 18th International Conference on Pattern Recognition (ICPR)*, 2006. Citado en página 39.

[153] **Zhang L., Ding Z., Li H., and Shen Y.** 3d ear identification based on sparse representation. *PLoS ONE*, 2014. Citado en página 56.

[154] **J. Lei, J. Zhou, and M. Abdel-Mottaleb**. Gender classification using automatically detected and aligned 3d ear range data. *In Biometrics (ICB), International Conference on*, 2013. Citado en las páginas 7 y 56.

[155] **L. Lu, Z. Xiaoxun, Z. Youdong, and J. Yunde**. Ear recognition based on statistical shape model. *In First International Conference on Innovative Computing, Information and Control*, 2006. Citado en las páginas 39 y 47.

[156] **Abdel-Mottaleb M. and J. Zhou**. Human ear recognition from face profile images. *Advances in Biometrics*, 2005. Citado en las páginas 37 y 39.

[157] **Alaraj M., Hou J., and Fukami. T**. A neural network based human identification framework using ear images. *TENCON*, 2010. Citado en las páginas 40 y 47.

[158] **Burge M. and Burger W.** Ear biometrics. *Springer*, 1998. Citado en página 39.

[159] **Choras M.** Image feature extraction methods for ear biometrics a survey. *In Computer Information Systems and Industrial Management Applications*, 2007. Citado en página 21.

[160] **Choras M.** Image pre-classification for biometrics identification systems. *Advances in Information Processing and Protection*, 2008. Citado en página 58.

[161] **Choras M.** Perspective methods of human identification: Ear biometrics. *Opto-Electronics Review*, 2008. Citado en las páginas 10, 39, y 44.

[162] **De Marsico M., Michele N, and Riccio D.** Hero: Human ear recognition against occlusions. *In IEEE Computer Society Conference on Computer Vision and Pattern Recog-nition Workshops (CVPRW)*, 2010. Citado en las páginas 40 y 42.

[163] **Grgic M., Delac K, and Grgic S.** Scface surveillance cameras face database. *Multimedia Tools Application*, 2011. Citado en las páginas 26 y 27.

[164] **Ibrahim M., Nixon M., and Mahmoodi S.** The effect of time on ear biometrics. *In International Joint Conference on Biometrics (IJCB)*, 2011. Citado en página 20.

[165] **Islam S. M., Davies R., Mian M., and Bennamoun A. S.** A fast and fully automatic ear recognition approach based on 3d local surface features. *In Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, 2008. Citado en página 52.

[166] **Jancosek M. and Pajdla T.** Multiview reconstruction preserving weakly supported surfaces. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[167] **Ozuysal M., Calonder M., Lepetit V., and Fua P.** Fast keypoint recognition using random ferns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2010.

[168] **Rahman M., Islam R., Bhuiyan N. I., Ahmed B., and Islam A.** Person identification using ear biometrics. *International Journal of The Computer, The Internet and Management*, 2007. Citado en página 39.

[169] **Yaqubi M., Faez K., and Motamed S.** Ear recognition using features inspired by visual cortex and support vector machine technique. *In International Conference on Computer and Communication Engineering (ICCCE)*, 2008. Citado en las páginas 40 y 49.

[170] **John Midgley**. Probabilistic eigenspace object recognition in the presence of occlusions. *Thesis, University of Toronto*, 2001. Citado en página 3.

[171] **K. Mohanapriya and M. Babu Me**. Ear recognition by feature extraction using force field transformation. *International Journal of Engineering and Computer Science*, 2015.

[172] **Boodoo-Jahangeer N. and Baichoo S.** Lbp-based ear recognition. *In Bioinformatics and Bioengineering*, 2013.

[173] **Dalal N.** Finding peple in images and videos. *PhD thesis, Institut National Polytechnique de Grenoble*, 2006.

[174] **Dalal N. and Triggs B.** Histograms of oriented gradients for human detection. *In Computer Vision and Pattern Recognition*, 2005.

[175] **Damer N. and Fuhrer B.** Ear recognition using multi-scale histogram of oriented gradients. *In Intelligent Information Hiding and Multimedia Signal Processing*, 2012.

[176] **Jamil N., AlMisreb A., and Halin A. A.** Illumination invariant ear authentication. *In Conference on Robot PRIDE - Medical and Rehabilitation Robotics and Instrumentation*, 2014. Citado en las páginas 56, 57, y 58.

[177] **Ratha N., Karu K., Chen S., and Jain A.** A real-time matching system for large fingerprint databases. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1996.

[178] **NY Daily News**. Mezuzah arsonist snagged by an ear thanks to facial recognition technology. *April*, 2013.

[179] **F.C.P.O. of Germany Police**. Crime statistics annual report. *Federal Criminal Police Office of Germany*, 2013.

[180] **Ibrahim Omaraa, Feng Lia, Hongzhi Zhanga, and Wangmeng Zuoa**. A novel geometric feature extraction method for ear recognition. *Expert Systems with Applications*, 2016.

[181] **Besl P. and McKey N.** A method for registration of 3d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1992.

[182] **Dollar P., Welinder P, and Perona P.** Cascaded pose regression. *In Computer Vision and Pattern Recognition*, 2010. Citado en página 14.

[183] **Hardeep P., Swadas P. B, and Joshi M.** A survey on techniques and challenges in image super resolution reconstruction. *International Journal of Computer Science and Mobile Computing*, 2013.

[184] **Lancaster P. and Salkauskas K.** Surfaces generated by moving least squares. *Mathematics of Computation*, 1981.

[185] **Promila P. and Laxmi V.** Palmprint matching using lbp. *In Computing Sciences (ICCS), International Conference on*, 2012.

[186] **Singh P. and Purkait R.** Observations of external earan indian study. *Journal of Comparative Human Biology*, 2009.

[187] **Torr P. and Zisserman A.** Robust computation and parametrization of multiple view relations. *In Computer Vision, Sixth International Conference on*, 1998.

[188] **Viola P. and Jones M. J.** Robust real-time face detection. *International Journal of Computer Vision*, 2004.

[189] **Viola P. and Jones M.** Rapid object detection using a boosted cascade of simple features. *In Computer Vision and Pattern Recognition (CVPR)*, 2001. Citado en página 13.

[190] **Wagner P.** Fisherfaces. *URL: http://www.bytefish.de/blog/fisherfaces/*, 2013. Citado en las páginas 114 y 115.

[191] **Yan P. and Bowyer K.** A fast algorithm for icp-based 3d shape biometrics. *In Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, 2005. Citado en las páginas 51, 52, y 57.

[192] **Yan P. and Bowyer K.** Biometric recognition using 3d ear shape. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2007. Citado en las páginas 24, 30, 34, 51, 52, y 58.

[193] **Yan P. and Bowyer K. W.** An automatic 3d ear recognition system. *In 3DPVT*, 2006. Citado en página 56.

[194] **Cappelli R., Ferrara M., and Maltoni D.** Minutia cylinder-code: A new representation and matching technique for fingerprint recognition. *Pattern Analysis and Machine Intelligence*, 2010.

[195] **Cappelli R., Ferrara M, and Maltoni D.** Fingerprint indexing based on minutia cylinder-code. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2011.

[196] **Fooprateepsiri R. and Kurutach W.** Ear based personal identification approach forensic science tasks. *Chiang Mai Journal of Science*, 2012. Citado en página 41.

[197] **Gadde R.** Phd thesis. *West Virginia University*, 2012.

[198] **Imhofer R.** Die bedeutung der ohrmuschel f Â¨ur die feststellung der identität. *Archiv für die Kriminologie*, 1906. Citado en las páginas 7 y 19.

[199] **Khorsandi R. and Abdel-Mottaleb M.** Gender classification using 2d ear images and sparse representation. *In Workshop on Applications of Computer Vision*, 2013.

[200] **Mukherjee R. and Ross A.** Indexing iris images. *In International Conference on Pattern Recognition*, 2008.

[201] **Mukundan R., Ong S., and Lee P.** Image analysis by tchebichef moments. *IEEE Transactions on Image Processing*, 2001. Citado en página 46.

[202] **Raghavendra R., Raja K., Pflug A., Yang B., and Busch C.** 3d face reconstruction and multimodal person identification from video captured using a smartphone camera. *In Proceedings of the 13th IEEE Conference on Technologies for Homeland Security (HST)*, 2013.

[203] **Rui Raposo, Edmundo Hoyle, Adolfo Peixinho, and Hugo Proença**. Ubear: A dataset of ear images captured on-the-move in uncontrolled conditions. *IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (SSCI-CIBIM)*, 2011. Citado en las páginas 28, 68, y 101.

[204] **Agarwal S., N. Snavely, I. Simon, and S. M. Seitzand R. Szeliski**. Building rome in a day. *IEEE 12th International Conference on Computer Vision*, 2009.

[205] **Ansari S. and Gupta P.** Localization of ear using outer helix curve of the ear. *International Conference on Computing: Theory and Applications*, 2007. Citado en las páginas 30 y 32.

[206] **Attarchi S., Faez K., and Rafiei A.** A new segmentation approach for ear recognition. *Advanced Concepts for Intelligent Vision Systems*, 2008. Citado en las páginas 30, 32, y 33.

[207] **Baker S., Matthews I., and Schneider J.** Automatic construction of active appearance models as an image coding problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2004.

[208] **Belongie S., Malik J, and Puzicha J.** Shape context: A new descriptor for shape matching and object recognition. *In Neural Information Processing Systems Conference (NIPS)*, 2000.

[209] **Billeb S., Rathgeb C., Buschbeck M. Reininger H., and Kasper K.** Efficient two-stage speaker identification based in universal background models. *Conference of the Biometrics Special Interest Group*, 2014.

[210] **Cadavid S. and Abdel-Mottaleb M.** 3d ear modeling and recognition from video sequences using shape from shading. *In 19th International Conference on Pattern Recognition (ICPR)*, 2008. Citado en página 52.

[211] **Cadavid S., Mahoor M., and Abdel-Mottaleb M.** Multi-modal biometric modeling and recognition of the human face and ear. *In IEEE International Workshop on Safety, Security Rescue Robotics (SSRR)*, 2009. Citado en las páginas 52, 53, y 61.

[212] **Islam S., Mian A. Bennamoun M., and Davies R.** A fully automatic approach for human recognition from profile images using 2d and 3d ear data. *In Proceedings of 3DPVT the Fourth Internatinoal Symposium on 3D Data Processing, Visualization and Transmission*, 2008. Citado en las páginas 52, 54, y 55.

[213] **Islam S., Bennamoun M., and Davies R.** Fast and fully automatic ear detection using cascaded adaboost. *In Applications of Computer Vision*, 2008. Citado en página 30.

[214] **Islam S., Davies R., Bennamoun M., and Mian A.** Efficient detection and recognition of 3d ears. *International Journal of Computer Vision*, 2011. Citado en las páginas 33, 53, y 55.

[215] **Islam S. M. S., Bennamoun M., Owens R., and Davies R.** Biometric approaches of 2d-3d ear and face: A survey. *Advances in Computer and Information Sciences and Engineering*, 2008. Citado en página 21.

[216] **Maity S. and Abdel-Mottaleb M.** 3d ear segmentation and classification through indexing. *IEEE Transactions on Information Forensics and Security*, 2014.

[217] **Mika S., Ratsch G., Weston J., Scholkopf B., and Muller K.** Fisher discriminant analysis with kernels. *In Neural Networks for Signal Processing IX Processing Society Workshop*, 1999.

[218] **Prakash S. and Gupta P.** An efficient ear recognition technique invariant to illumination and pose. *Telecommunication Systems Journal, special issue on Signal Processing Applications in Human Computer Interaction*, 2011. Citado en las páginas 26, 40, y 45.

[219] **Prakash S. and Gupta P.** An efficient technique for ear detection in 3d: Invariant to rotation and scale. *In The 5th IAPR International Conference on Biometrics (ICB)*, 2012. Citado en las páginas 31, 32, y 33.

[220] **Prakash S. and Gupta P.** An effient ear localization technique. *Image and Vision Computing*, 2012. Citado en las páginas 30 y 33.

[221] **Prakash S. and Gupta P.** An efficient ear recognition technique invariant to illumination and pose. *Telecommunication Systems Journal*, 2013. Citado en página 58.

[222] **Wang S.** An improved normalization method for ear feature extraction. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 2013.

[223] **Thomas B. Sebastian, Philip N. Klein, and Benjamin B. Kimia.** Shock-based indexing into large shape databases. *In European Conference on Computer Vision*, 2002. Citado en página 5.

[224] **Pamplona M. Segundo, Silva L., and P. Bellon O. R.** Improving 3d face reconstruction from a single image using half frontal face poses. *In 19th IEEE Conference on Image Processing (ICIP)*, 2012.

[225] **Joginder Singh**. Ear pattern recognition and compression. *International Journal of Computer Science and Communication*, 2015.

[226] **Santosh H. Suryawanshi**. The ear as a biometric. *International Journal of Computer Science and Mobile Computing*, 2015.

[227] **Ahonen T. and A. Hadidand M. Pietikainen**. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence*, 2006. Citado en página 13.

[228] **Ahonen T., E. Rahtu, and V. Ojansivuand J. Heikkila**. Recognition of blurred faces using local phase quantization. *International Conference on Pattern Recognition*, 2008.

[229] **Hara T., Kubo H., Maejima A., and Morishima S.** Fast accurate 3d face model generation using a single video camera. *In Pattern Recognition (ICPR)*, 2012.

[230] **Ojala T., M. Pietikainen, and T. Maenpaa**. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2002.

[231] **Vikram T., Chidananda Gowda K., Guru D., and Urs S.** Face indexing and retrieval by spatial similarity. *In Image and Signal Processing*, 2008.

[232] **Yuizono T., Wang Y., Satoh K., and Nakayama S.** Study on individual recognition for ear images by using genetic local search. *In Proceedings of the 2002 Congress on Processing Scociety of Japan (IPSJ) Kyushu Chapter Symposium*, 2002. Citado en las páginas 39 y 49.

[233] **Park U. and Jain A.** 3d model-based face recognition in video. *Advances in Biometrics*, 2007.

[234] **Cantoni V., Dimov D., and Nikolov A.** 3d ear analysis by an egi representation. *Biometric Authentication: Lecture Notes in Computer Science*, 2014. Citado en página 60.

[235] **Iannarelli A. V.** Ear identification. *Paramount Publishing Company*, 1964. Citado en las páginas 7, 18, 20, y 64.

[236] **Ojansivu V. and Heikkil J.** Blur insensitive texture classification using local phase quantization. *In Image and Signal Processing*, 2008.

[237] **Struc V. and Pavesic N.** The complete gabor-fisher classifier for robust face recognition. *EURASIP J. Adv. Signal Process*, 2010.

[238] **Hayward W., Tarr M, and Corderoy A.** Recognizing silhouettes and shaded images across depth rotations. *Perception*, 1999. Citado en página 4.

[239] **Zhao W., Chellappa R., Phillips P. J., and Rosenfeld A.** Face recognition: A literature survey. *ACM Comput. Survey*, 2003.

[240] **Burgos-Artizzu X., Perona P., and Dollár P.** Robust face landmark estimation under occlusion. *ICCV*, 2013.

[241] **He X. and Yung N.** Corner detector based on global and local curvature properties. *Optical Engineering*, 2008.

[242] **Wang X. and Yuan W.** Human ear recognition based on block segmentation. *In International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 2009. Citado en las páginas 40 y 43.

[243] **Wang X. and Yuan W.** Gabor wavelets and general discriminant analysis for ear recogniton. *In 8th World Congress on Intelligent Control and Automation (WCICA)*, 2010. Citado en las páginas 38 y 40.

[244] **Zhu X. and Ramanan D.** Face detection, pose estimation and landmark localization in the wild. *In Computer Vision and Pattern Recognition*, 2012. Citado en página 56.

[245] **Wang X.-q., Xia H.-y., and Wang Z.-l.** The research of ear identification based on improved algorithm of moment invariants. *In Third International Conference on Information and Computing (ICIC)*, 2010. Citado en las páginas 40 y 42.

[246] **Furukawa Y. and Ponce J.** Accurate, denseand robust multiview stereopsis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2010.

[247] **Guo Y. and Xu Z. i. W.** Ear recognition using a new local matching approach. *In International Conference on Image Processing*, 2008. Citado en página 40.

[248] **Liu Y., Zhang B., and Zhang D.** Ear-parotic face angle: A unique feature for 3d ear recognition. *Pattern Recognition Letters*, 2014. Citado en página 60.

[249] **Wang Y., Chun Mu Z., and Zeng H.** Block based and multi resolition methods for ear recognition using walelste transform and uniform local binary patterns. *In 19th International Conference on Pattern Recognition (ICPR)*, 2008. Citado en página 40.

[250] **Xu Y. and Zeng W.** Ear recognition based on centroid and spindle. *International Workshop on Information and Electronics Engineering*, 2012. Citado en página 58.

[251] **Mu Z., Yuan L., Xu Z., Xi D., and Qi S.** Shape and structural feature based ear recognition. *Advances in Biometric Person Authentication*, 2005. Citado en página 39.

[252] **Xie Z. and Mu Z.** Ear recognition using lle and idlle algorithm. *In 19th International Conference on Pattern Recognition (ICPR)*, 2008. Citado en las páginas 40 y 48.

[253] **Wang Z.-q. and Yan X.-d.** Multi-scale feature extraction algorithm of ear image. *In International Conference on Electric Information and Control Engineering (ICEICE)*, 2011. Citado en página 40.

Anexos

A

Estudio de oclusión de orejas

A.1. Resultado

Entre septiembre 2013 y agosto 2014, se realizó en la Universidad de Salamanca un estudio básico sobre la posibilidad de observar las orejas de las personas con o sin oclusión. El objetivo era conseguir una sencilla estimación sobre si era posible obtener una visión clara de la oreja en un escenario real.

Durante el tiempo del estudio, se contó el número de personas cuyas orejas estaban ocluidas, mientras caminaban para acceder a la facultad de ciencias y/o la facultad de derecho del campus de la Universidad. Adicionalmente, también se tomaron notas sobre el género y las condiciones climáticas exteriores. En total, se observaron 5.431 personas. Las observaciones se realizaron sólo durante ciertos días de cada mes en horario de 9:00 AM a 3:00 PM., el estudio fue totalmente empírico y sólo buscaba estimar cuan probable es tener una visión clara de la oreja de un individuo en el mundo real; es altamente probable que el horario influya en la forma en cómo se visten muchos sujetos. Finalmente, el número total de personas que se observan dentro de un determinado período de tiempo fue menor en los meses que corresponden a las vacaciones de verano. Los datos se clasificaron entre hombres y mujeres y el tipo de oclusión, se definieron seis categorías, las que se detallan a continuación.

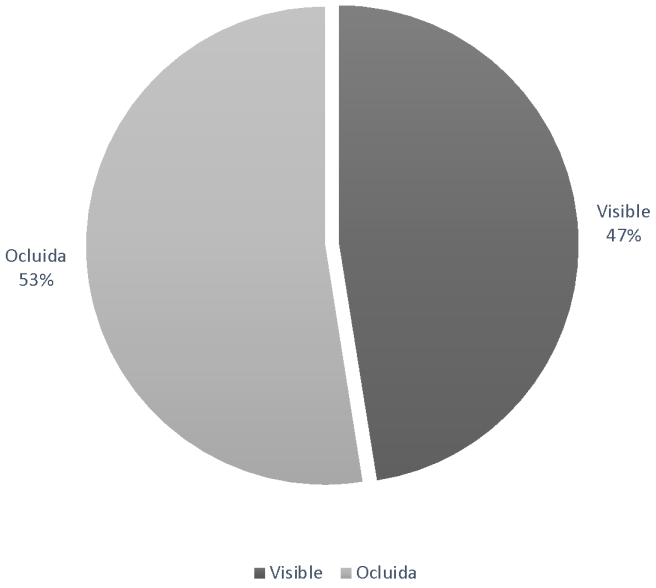


Figura A.1.1: Oclusión de las orejas de todos los sujetos del estudio

- **Parcialmente ocluida:** al menos una tercera parte de la oreja se encuentra ocluida por el cabello.
- **Completamente ocluida:** la oreja está completamente ocluida por el cabello.
- **Accesorios:** la oreja se ocluye por una gorra, u otro accesorio en la cabeza.
- **Pendientes grandes:** la persona lleva pendientes u otra decoración grande que puede que ocluya completamente el lóbulo u otras partes de la oreja.
- **Pendientes colgantes:** la persona lleva pendientes que se adjuntan al lóbulo; estos pendientes pueden producir errores en el proceso de segmentación.
- **Audífonos:** la persona está usando auriculares que pueden ocluir la concha u ocluir completamente la oreja.

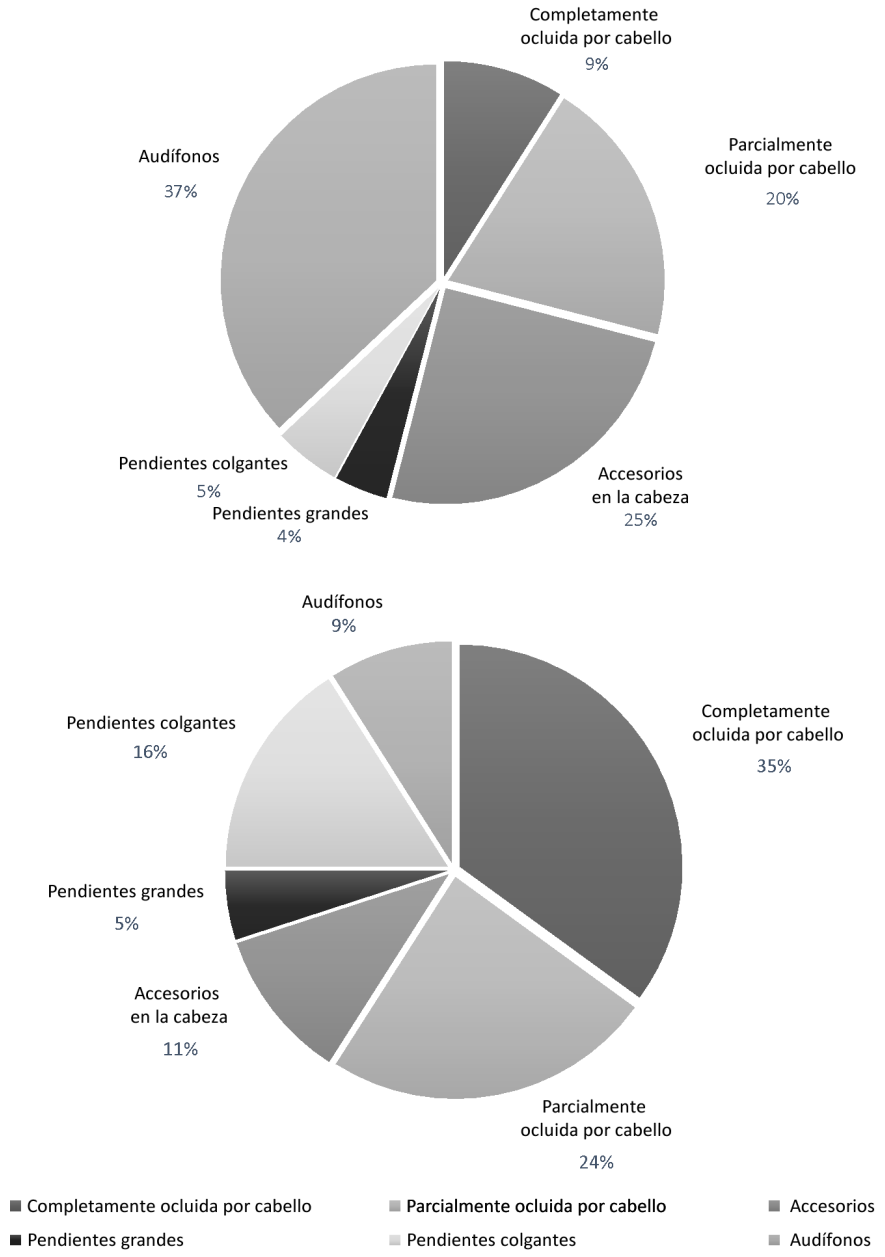


Figura A.1.2: Tipos de oclusión en hombres (gráfico superior) y en mujeres (gráfico inferior)

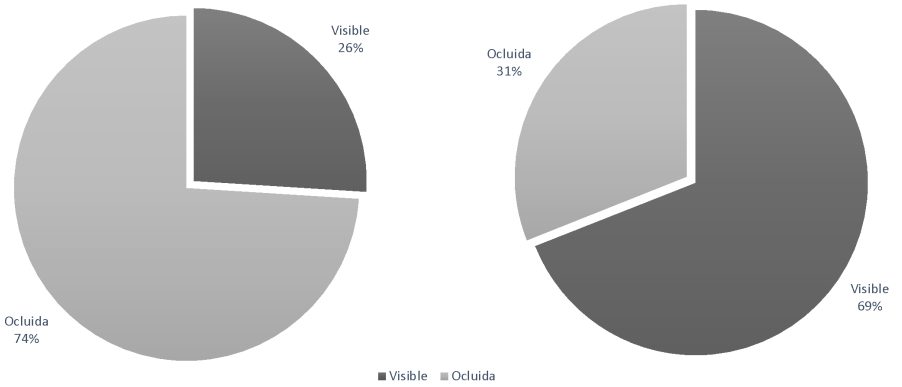


Figura A.1.3: Oclusión de las orejas para todas las personas en el estudio, mujeres (izquierda), hombres (derecha)

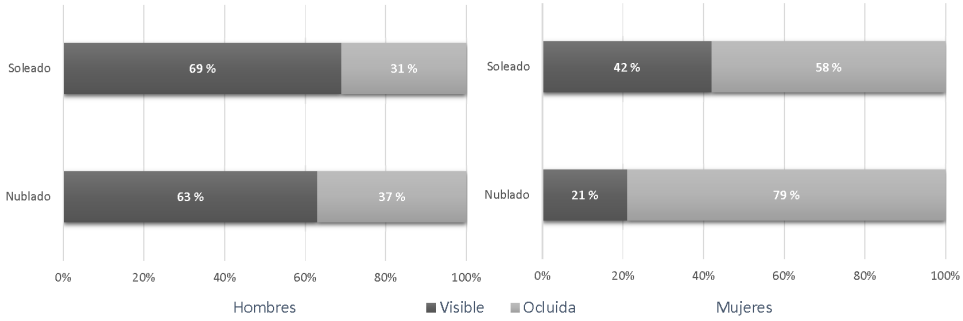


Figura A.1.4: Impacto de las condiciones climatológicas en la oclusión de la oreja

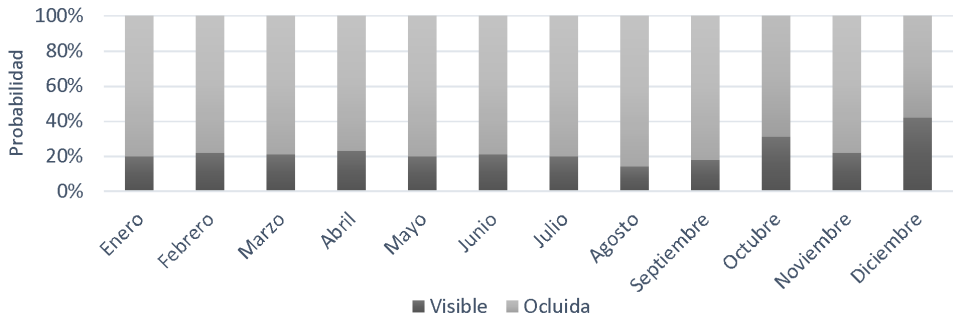


Figura A.1.5: Probabilidad de observar la oreja ocluida (mujeres)

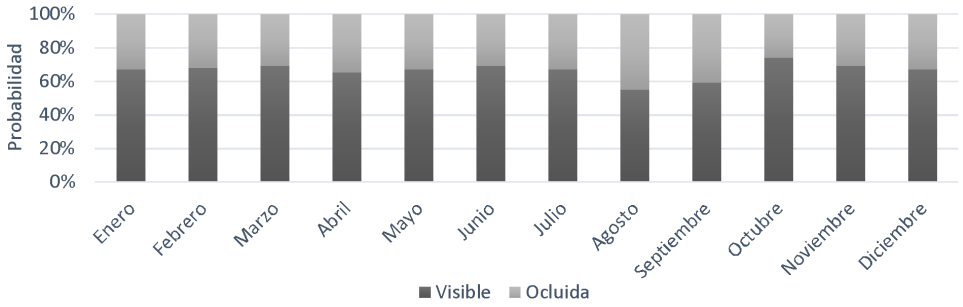


Figura A.1.6: Probabilidad de observar la oreja ocluida (hombres)

A.2. Conclusión

Este estudio presenta una visión general sobre la probabilidad media de obtener imágenes visibles de la oreja en lugares públicos. Debemos ser conscientes que los resultados aquí planteados contiene algún tipo de sesgo debido a varios factores, tales como el lugar y la hora del día. Observamos que la probabilidad de que la oreja esté visible es mucho mayor para los hombres que para las mujeres. La visibilidad cambia ligeramente debido a las condiciones meteorológicas.

Es importante recordar que el número total de observaciones es relativamente pequeño y que se requieren más datos para tener una mejor impresión sobre la visibilidad de la oreja en público. También nos limitamos a un sólo lugar, lo que sin duda tiene un impacto en nuestros resultados. Es probable que el uso de audífonos pueda variar en un lugar donde la gente tiene que interactuar con los demás. El número de personas con gorras y bufandas es probable que sea menor dentro de un edificio por ejemplo, un centro comercial.

B

Estándares y términos comunes en biometría

B.1. Sistema biométrico genérico

Los componentes generales y el flujo de trabajo de un sistema biométrico se describe en la normativa estándar ISO/IEC 2382-37: 2012 [1]. En esta investigación, en particular nos centramos en las etapas de captura, pre-procesamiento, el algoritmo de comparación y concordancia y el subsistema de decisión.

Evaluamos los sistemas de reconocimiento de orejas en dos modos diferentes, que son la verificación y el modo de identificación. La verificación se refiere a un modo operativo, donde las imágenes presentes (también conocidas como imágenes a investigar o explorar) se asocian con una pretención de identidad. El sistema verifica o deniega esta afirmación, lo que significa que toma una decisión binaria sobre si o no, la afirmación es verdadera. En el modo de identificación, solo se tiene una imagen exploratoria de prueba sin una afirmación de identidad. El sistema retornará la identidad del candidato más probable al que pertenece esta imagen, con base a las imágenes de la base de datos.

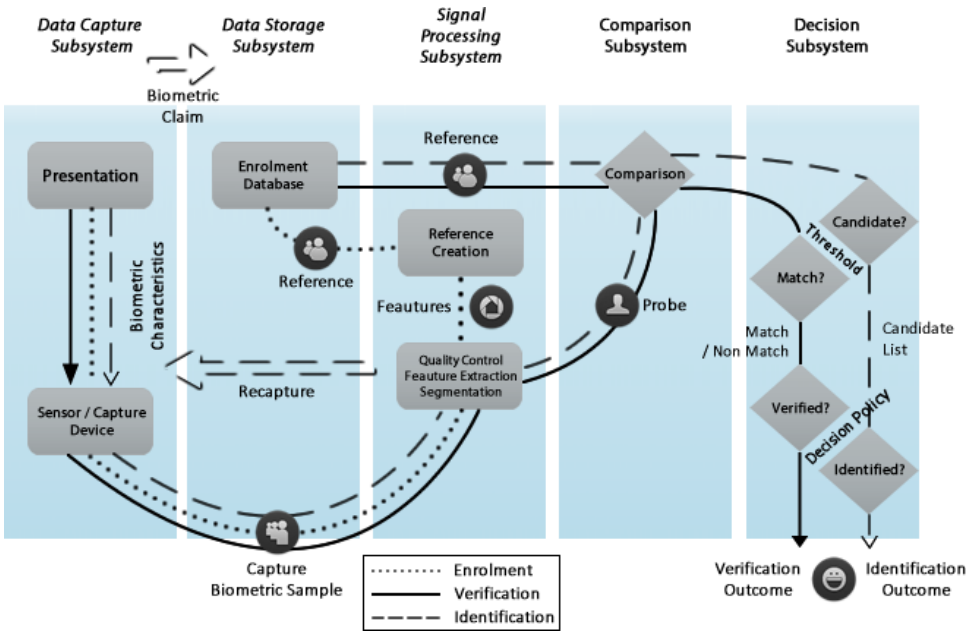


Figura B.1.1: Sistema biométrico genérico según lo definido en la norma estándar ISO/IEC SC37 SD11

B.2. Vocabulario armonizado

La norma ISO-IEC [47] define una serie de términos que están bien establecidos en la comunidad investigadora en el área de la biometría. En la presente investigación se utilizan múltiples términos todos ellos basados en los componentes de un sistema biométrico genérico, como se describe en la figura B.1.1.

B.2.1. Términos generales

Al describir los sistemas biométricos, hacemos uso de una serie de conceptos estandarizados. Estos términos se definen en la norma estándar ISO/IEC JTC SC37 SD11 [47]. Para una descripción completa del vocabulario utilizado en biometría, el lector puede referirse al documento original estándar. Los siguientes son conceptos importantes a destacar:

- **Imagen a investigar - sondear/imagen exploratoria:** una muestra o conjunto de características biométricas de entrada para un algoritmo que lo utilizará como el objeto de comparación a una referencia(s) biométrica.

- **Referencia:** una o más muestras almacenadas, plantillas o modelos biométricos atribuidos a datos biométricos de un sujeto y que se utilizan como el objeto contra el que se realizará para la comparación biométrica.
- **Características:** números o etiquetas extraídas de muestras biométricas y utilizadas para la comparación.
- **Plantillas:** un conjunto de características biométricas almacenadas que sirven de base y pueden ser directamente comparadas con un objeto o características a investigar.
- **Enrolamiento:** es el acto de crear y almacenar un registro de datos de inscripciones biométricas de acuerdo con una política de inscripción.
- **Comparación:** la estimación, cálculo o medición de similitud o disimilitud entre la imagen(s) de prueba (exploración) y la referencia(s) biométrica.

B.2.2. Rendimiento y fallos del sistema

Al describir el rendimiento de un sistema biométrico, la norma propone los siguientes indicadores de desempeño. Todas estas definiciones se han tomado del estándar *ISO/IEC JTC1 SC37 Biometrics* [47].

- ***Failure to Enrol (FTE):*** proporción de la población para quienes el sistema falla al completar el proceso de inscripción. Estos fallos pueden ocurrir en cualquier momento durante el proceso de captura, proceso de segmentación o el proceso de extracción de características.
- ***False Match Rate (FMR):*** proporción de intentos falsamente declarados como coincidencia en la comparación de plantillas distintas a la del sujeto buscado. Es importante resaltar que este indicador de rendimiento solamente mide el rendimiento de los algoritmos y no el rendimiento del sistema.
- ***False Non-Match Rate (FNMR):*** proporción de muestras declaradas falsamente como no coincidencia con la plantilla del mismo usuario. Al igual que el anterior este indicador sólo mide el rendimiento de los algoritmos y no el rendimiento del sistema.

- ***False Accept Rate (FAR):*** proporción de los intentos de verificación de identidades falsas que han sido incorrectamente confirmadas. Este indicador de rendimiento se refiere al sistema completo, lo que significa que incorpora el FTE.

- ***False Reject Rate (FRR):*** proporción de los intentos de verificación de identidades correctas que han sido incorrectamente denegadas. Este indicador de rendimiento se refiere al sistema completo, lo que significa que también incorpora el FTE.

- ***Equal Error Rate (EER):*** un punto operacional en el que el FAR y FRR son iguales. A menudo utilizado para reportar el desempeño en el modo de verificación.

- ***Penetration Rate (PEN):*** medida del número promedio de plantillas preseleccionadas como una fracción del número total de plantillas.

- ***Rank-n Identification Rate (R1 or IR):*** tasa de identificación de los intentos de identificación de los usuarios inscritos en el sistema en el que el identificador correcto del usuario se encuentra entre quienes regresaron para reportar el desempeño en el modo de verificación. Dónde n es el rango máximo de los verdaderos positivos en una lista ordenada de candidatos.

- ***Preselection Error (PSE):*** error que ocurre cuando la correspondiente plantilla de registro no se encuentra en el subconjunto de candidatos a pesar de que se da una muestra de las mismas características para el mismo usuario.

C

Laboratorio de procesamiento de imágenes

El presente apéndice describe de forma general los distintos componentes del sistema. La investigación ha girado entorno a los algoritmos utilizados para el proceso de detección y reconocimiento, sin embargo es importante resaltar que alrededor de ese núcleo se ha desarrollado una herramienta de software, la cual ha sido nombrada *Laboratorio para el procesamiento de imágenes* ya que nos permite realizar tareas de normalización, transformación e incluso aplicar los algoritmos de detección y reconocimiento descritos en el presente documento.

C.1. Interfaz del sistema

El escritorio de trabajo que recibe al usuario al ejecutar por primera vez la aplicación esta construído con un concepto multimodal, donde el sujeto puede trabajar con múltiples ventanas a la vez, contenidas en un entorno único de trabajo. La figura C.1.1 muestra la interfaz principal, la cual presenta un área de trabajo central donde se aprecian las imágenes, un conjunto de menús desplegables en la parte superior, y una sección de herramientas con íconos de acceso rápido a tareas frecuentes.

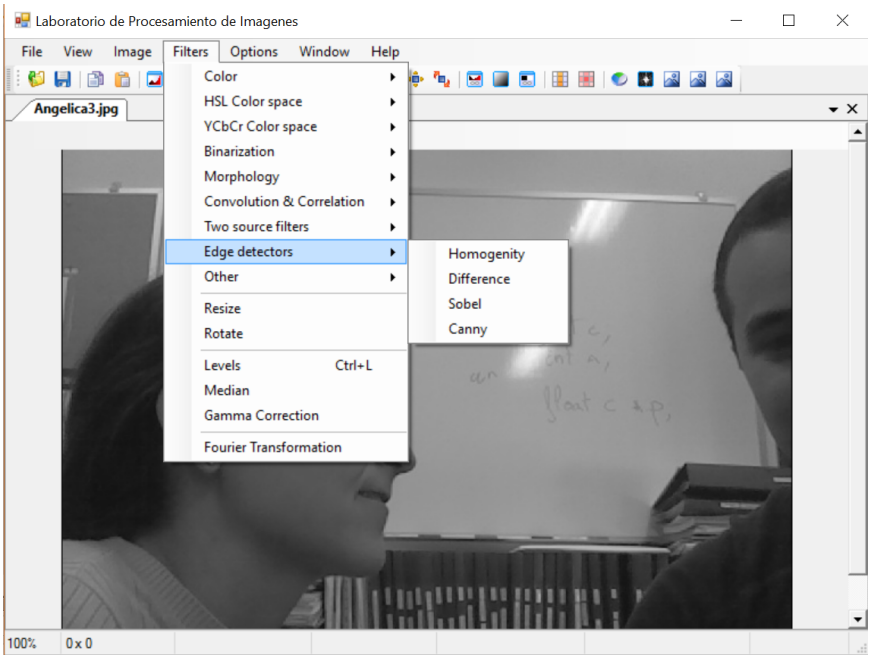


Figura C.1.1: Interfaz principal del sistema

Los menús se dividen en siete grupos; las tareas principales se concentran en los menús de filtros (*filters*) e imagen (*image*) los que contienen los algoritmos de pre-procesamiento, por ejemplo Canny o Sobel para detectar bordes y las opciones básicas como ser recortar, acercar, alejar y clonar una imagen.

La experiencia del usuario es un aspecto de vital importancia y, en cierta medida, determina el grado de aceptación de la herramienta por parte de los usuarios, el aspecto de presentación y las opciones de interacción son altamente personalizables ya que el usuario puede ajustar la posición de las barras de herramientas, el tamaño y posición de las ventanas, así como el fondo del espacio de trabajo. La figura C.1.2 muestra como el usuario trabaja con dos imágenes a la vez.

El usuario a este nivel puede decidir trabajar con N cantidad de imágenes, aplicarles algoritmos de detección de bordes, modificar colores, visualizar el histograma, cambiar el tamaño o recortar la imagen e incluso aplicar la transformada de Fourier. Todas son tareas posibles dentro del entorno de trabajo.

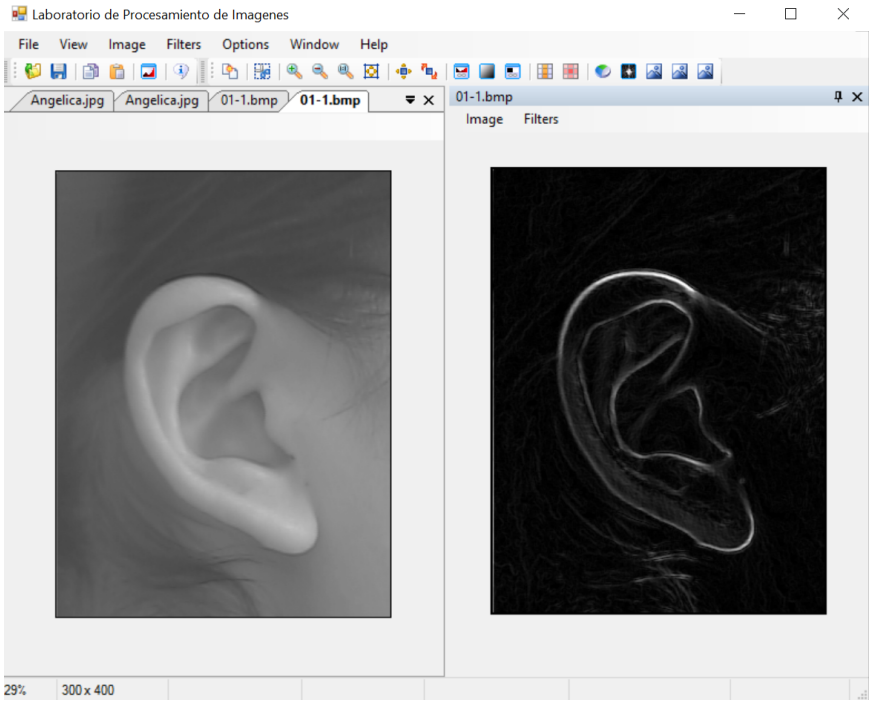


Figura C.1.2: Editando dos imágenes a la vez en el sistema

En este punto, podemos afirmar que la herramienta desarrollada es un software editor de imágenes que puede ser utilizado tanto a nivel de laboratorio como en actividades cotidianas; el aplicativo soporta formatos de imagen estándar como el BMP, JPEG y PNG.

Finalmente, el usuario es capaz de realizar tareas básicas de reconocimiento y detección a partir de imágenes o vídeo. En la figura C.1.3 se aprecia el proceso de detección y reconocimiento de orejas en funcionamiento; el usuario reconocido es etiquetado en la parte inferior de la pantalla. La figura C.1.4 muestra ejemplos del resultado de aplicar los algoritmos descritos en el presente documento dentro del sistema desarrollado aplicable a cualquier imagen y por lo tanto servirá para otras posibles investigaciones.

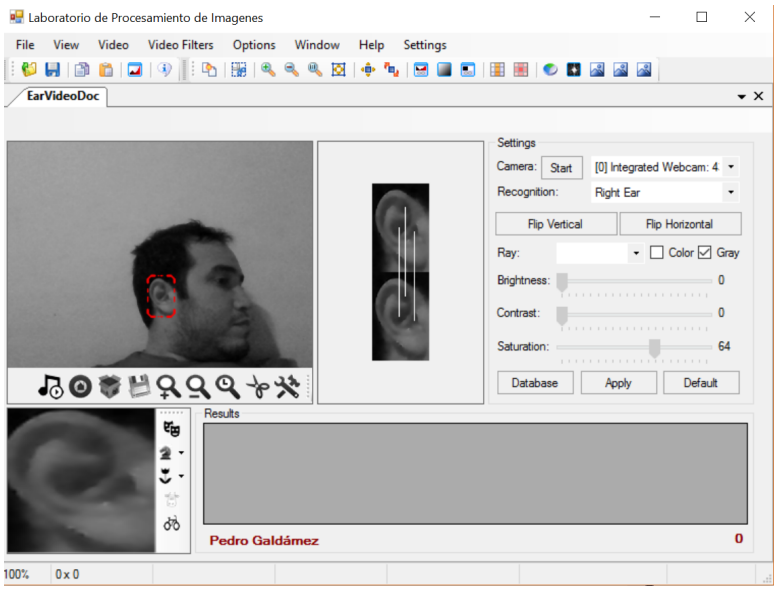
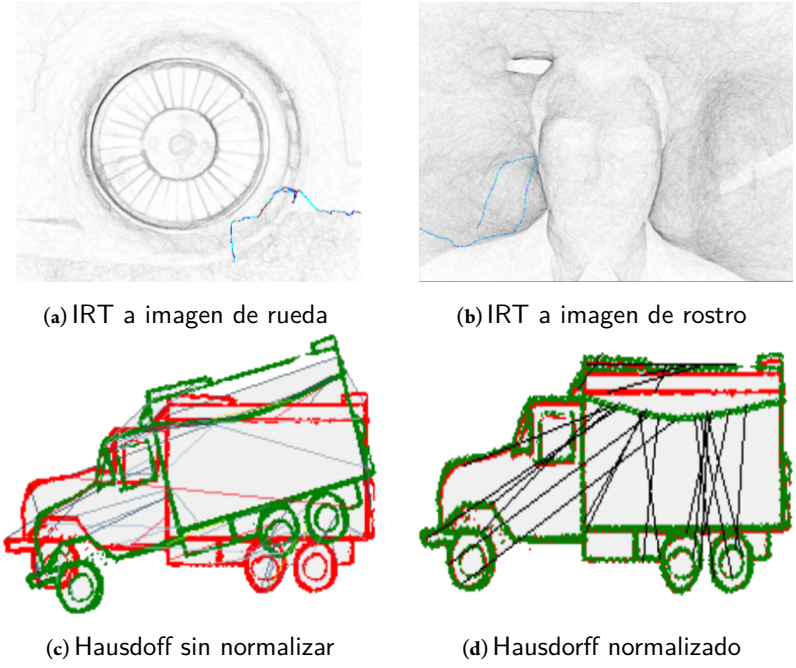


Figura C.1.3: Reconocimiento de orejas en el sistema



(a) IRT a imagen de rueda

(b) IRT a imagen de rostro

(c) Hausdoff sin normalizar

(d) Hausdorff normalizado

Figura C.1.4: Ejemplo de los algoritmos aplicables en el sistema