DESARROLLO DE UN ALGORITMO DE RE-IDENTIFICACIÓN MULTI-MODAL DE PERSONAS PARA MEJORAR LA ASISTENCIA PERSONALIZADA EN UNA CASA FAMILIAR

BRYAN STEVEN BETANCUR SANCHEZ
LINA ALEJANDRA PLAZAS PIRABÁN

UNIVERSIDAD SANTO TOMÁS
INGENIERÍA ELECTRÓNICA
BOGOTÁ D.C.
2021

DESARROLLO DE UN ALGORITMO DE RE-IDENTIFICACIÓN MULTI-MODAL DE PERSONAS PARA MEJORAR LA ASISTENCIA PERSONALIZADA EN UNA CASA FAMILIAR

BRYAN STEVEN BETANCUR SANCHEZ
LINA ALEJANDRA PLAZAS PIRABÁN

Proyecto de grado presentado como requisito para optar al título de
INGENIERO ELECTRÓNICO

UNIVERSIDAD SANTO TOMÁS
INGENIERIA ELECTRÓNICA
BOGOTÁ D.C.
2021

AGRADECIMIENTOS

En primer lugar a nuestros padres por el apoyo incondicional que nos brindaron en todo momento, porque en los momentos más difíciles creyeron en nosotros y nos impulsaron a ser mejores, ellos son la razón principal por la cual logramos culminar tanto el proyecto de grado como la carrera con éxito.

De igual manera a cada uno de los docentes que con paciencia y mucha dedicación nos compartieron sus conocimientos, ayudándonos cada día no solo a ser mejores profesionales sino también brindándonos una educación integral.

Índice general

Índice de tablas

Índice de figuras

Glosario

- **Algoritmo Multimodal**: integración de múltiples algoritmos como lo son en este caso reconocimiento facial, por voz y por características soft-biométricas en un solo desarrollo.

- **Aprendizaje automático**: es una rama de la Inteligencia Artificial, en la cual se busca desarrollar algoritmos que a partir de datos ejemplo, logren resolver nuevos problemas [1].

- **Asistencia robótica personalizada**: se refiere a que el robot con base en una previa re-identificación de la persona, pueda brindar servicios dependiendo del humano con el que esté interactuando [2].

- **Características biométricas**: son aquellos rasgos fisiológicos que hacen único a un ser humano, por ejemplo el rostro, la huella dactilar, propiedades de la voz, características de la retina, etc [3].

- **Características soft-biométricas**: son propiedades físicas de los seres humanos que proporcionan información relevante pero no distintiva para una correcta identificación, como por ejemplo: color de piel, color de cabello, color de ojos, etc [4].

- **Inteligencia Artificial**: es la combinación de algoritmos que tienen como objetivo aprender comportamientos de los seres humanos [5].

- **Matriz de confusión**: es un conjunto de métricas que permiten evaluar el desempeño de algoritmos de aprendizaje automático.

- **Re-identificación de personas**: es el proceso en el cual se asocian imágenes o audios a un mismo individuo a pesar de que la imagen o audio se obtenga desde diferentes cámaras y/o micrófonos en distintas ocasiones [6].
- **Robótica social**: es una rama de la robótica, la cual se encarga de que la interacción humano-robot sea lo más fluida posible [7].

Resumen

Este documento presenta el desarrollo de un algoritmo de re-identificación multimodal para mejorar la interacción Humano-Robot en el ámbito de asistencia doméstica. De esta manera, se integraron diferentes estrategias de reconocimiento de personas como lo son reconocimiento facial, por voz y por características soft-biométricas (color de cabello, ojos y piel). Para esto, en primer lugar se realizó una consulta bibliográfica donde se eligieron posibles algoritmos a utilizar; luego se implementaron y se realizaron diferentes pruebas con el fin de elegir los algoritmos que presentaban mejores resultados por cada estrategia de re-identificación, después se integraron en un único desarrollo basado en regresión lineal múltiple el cual tuvo un porcentaje de acierto del 97.4 %. De igual manera, se implementó todo el sistema en ROS (sistema operativo robótico) y se realizaron pruebas donde se evaluó si el algoritmo reconocía órdenes básicas personalizadas.

Introducción

En los últimos años, el estudio de la Inteligencia Artificial (IA) ha tomado gran relevancia debido a sus múltiples aplicaciones en la vida cotidiana, es así como abarca una gran cantidad de aplicaciones y están orientados a satisfacer diferentes necesidades. Dentro de las técnicas con mayor uso se encuentra el aprendizaje automático, que consiste en que una máquina pueda aprender por sí misma; esto hace referencia al desarrollo de programas informáticos que pueden generalizar cuando se exponen a nuevos datos. [1].

Asimismo, unas de las aplicaciones más importantes del aprendizaje automático son los sistemas de reconocimiento de personas, ya sea reconocimiento facial o por voz. Estos algoritmos son de gran ayuda en actividades de rescate, sistemas de seguridad como por ejemplo el evitar fraude y vandalismos en cajeros automáticos, la asistencia doméstica, entre otros [8].

A pesar de los múltiples avances en este campo, el reconocimiento de personas continúa siendo un reto, puesto que al implementar una estrategia en específico, se pueden presentar varios inconvenientes, por ejemplo: el desempeño de un algoritmo de reconocimiento facial se puede ver afectado en personas con rostros muy similares o por factores externos como la iluminación en el ambiente [9], de igual manera ocurre con el reconocimiento de voz que puede presentar fallas en entornos con mucho ruido [10]. Por lo anterior, el presente proyecto pretende desarrollar un

algoritmo que integre distintas estrategias de identificación como reconocimiento facial, por voz y por características soft-biométricas (género, color de cabello, de ojos y de piel). Asimismo, se enfoca en el ámbito doméstico, con el fin de que al haber un buen reconocimiento, la interacción humano-robot mejore y se le pueda brindar al usuario una experiencia más personalizada.

De acuerdo a lo anterior, el presente documento inicia exponiendo el problema observado, luego presenta otros trabajos realizados con relación a cada una de las estrategias de identificación, posteriormente se expone la importancia del proyecto y los conceptos teóricos sobre los cuales se basa este trabajo. Después se muestra la metodología utilizada, así como las diferentes pruebas y evaluaciones que se realizaron para elegir los algoritmos con mejor desempeño, más adelante se explica la integración de los algoritmos elegidos en un único desarrollo así como la implementación del sistema en ROS y finalmente se presentan las correspondientes conclusiones y los posibles trabajos futuros que pueden surgir a partir del proyecto.

Capítulo 1

Planteamiento del problema

La robótica social es una rama de la ingeniería que se encarga del estudio de los robots diseñados para lograr una interacción con los seres humanos y el ambiente que los rodea [7]. Esta rama de la robótica tiene como objetivo mejorar la calidad de vida de las personas por medio de robots que adquieren comportamientos sociales acordes al entorno en el que se encuentran. Para lograr cumplir este objetivo es primordial una correcta re-identificación de personas, la cual le permita al robot reconocer y memorizar a un humano de forma eficiente[11]. Lo mencionado anteriormente se puede utilizar en una amplia gama de escenarios del mundo real, por ejemplo: tareas domésticas, realizar compras, ofrecer asistencia, brindar orientación, habilidad para entretener, proporcionar vigilancia y actividades de rescate [12] [13] [14], etc.

Pese a los diferentes desarrollos alcanzados por la robótica social, la re-identificación continúa siendo un gran reto en el ámbito científico y académico, debido a que se dificulta implementar este tipo de técnicas en entornos no estructurados en condiciones ambientales cambiantes, y con una estrecha interacción humano-robot; por ejemplo, en entornos domésticos. Debido a lo anterior, actualmente la competencia mundial *RoboCup@Home* [15] tiene como objetivo incentivar el desarrollo de tecnologías de servicio y asistencia de robots con alta relevancia para

futuras aplicaciones domésticas y/o personales.

Por otro lado, si bien durante los últimos años se han realizado importantes avances en la identificación de personas, los mismos están basados en técnicas y estrategias particulares tales como identificación facial [9][16][17], identificación de voz [18][19], identificación por rasgos físicos notables [20], entre otros; por lo cual se plantea la siguiente problemática: ¿Es posible mejorar la re-identificación de personas y con ello la captación de órdenes para asistencia personalizada en un entorno familiar, por medio de un algoritmo que integre distintas estrategias tales como reconocimiento facial, identificación de voz y de características soft-biométricas(color de piel, color de cabello, color de ojos)?

Capítulo 2

Antecedentes

En los últimos años la re-identificación de personas ha tomado mayor relevancia en el marco HRI *(Human - Robot interaction)*, puesto que esta le permite al usuario tener una experiencia personalizada con los robots. Debido a lo anterior, se han planteado diversas técnicas y algoritmos para la correcta identificación facial, identificación de voz e identificación por características soft-biométricas. A continuación, serán mencionados algunos ejemplos encontrados en las bases de datos IEEE, *ScienceDirect* y Google Académico.

El artículo [16] se centra en el reconocimiento de la persona por medio de la identificación facial y propone un algoritmo donde se integran técnicas tales como: *Eigenface*, la cual busca que la cara que se desea analizar se vea aumentada, y de este modo lograr reconocer detalles y rasgos físicos destacados con mayor facilidad. El algoritmo *Eigenupper* el cual analiza la expresión del rostro centrandose en la región de la boca y el algoritmo*EigenTzone* que analiza únicamente los ojos y la nariz, puesto que estos contienen información destacable de los seres humanos. Dichos métodos se utilizaron para crear un sistema de seguridad multiusuario en un entorno de trabajo computacional, el cual provee a cada usuario únicamente los archivos a los que tiene acceso.

De igual manera en los artículos [9][17] se utiliza la técnica de aprendizaje profundo para la identificación facial, mostrando buenos resultados, como se muestra en la figura 2.1:

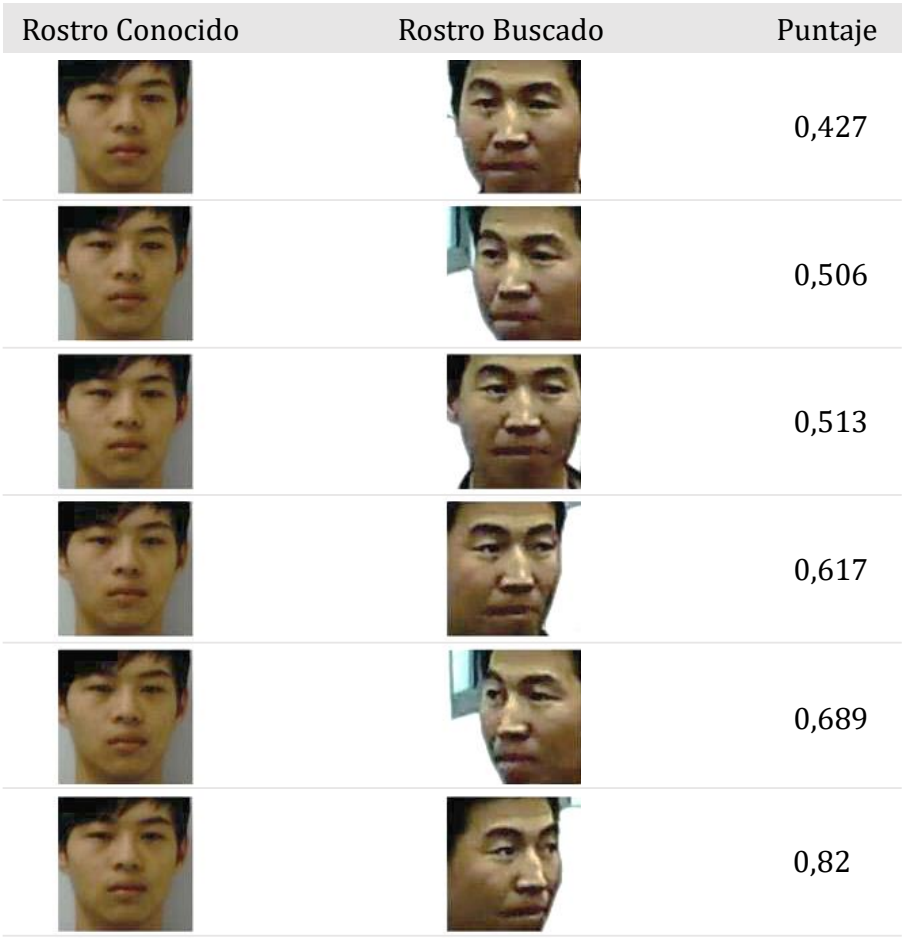

Figura 2.1. Resultados reconocimiento facial por aprendizaje de refuerzo [19]

En la figura 2.1 se plantea la situación de un rostro que necesita ser identificado, para esto se realiza una comparación con imágenes de caras ya conocidas. Los resultados se dan en un rango de 0 a 1 donde 0 representa coincidencia total y 1 gran diferencia entre los rostros, lo cual quiere decir que en este caso el rostro a identificar tiene una mayor

similitud con el sujeto mostrado en la comparación 1. Por otra parte, cabe destacar que ambos artículos llegan a una conclusión similar, a pesar de que el reconocimiento facial demuestra ser bastante eficiente, este no serviría por sí solo para realizar una identificación exhaustiva de personas, ya que en el caso de gemelos o personas muy parecidas los resultados que arrojaría serían muy similares, lo cual no permitiría realizar una identificación correcta.

De igual manera uno de los algoritmos más recientes para reconocimiento facial se encuentra en el artículo [21], en el cual presentan un nuevo modelo IE-CNN *(Internal and External features Convolutional Neural Network)* en donde se mejoran las características internas y externas de la cara para lograr aumentar los falsos positivos y de está manera mejorar la precisión de reconocimiento.

Por otro lado, en cuanto al reconocimiento de voz se debe tener en cuenta dos factores: el primero es el reconocimiento de palabras y el segundo la identificación de personas por diferencia de voces. El primer factor se ha trabajado en los artículos [18][19] los cuales utilizan MFCC *(Mel Frequency Cepstral Coefficients)* que en primer lugar toma muestras de voz como entrada, después procesa la información realizando transformada rápida de Fourier, luego se pasa la señal por un banco de filtros Mel y finalmente se aplica la transformada discreta de Fourier para obtener los coeficientes de Mel. En la figura 2.2, se muestra gráficamente el algoritmo:

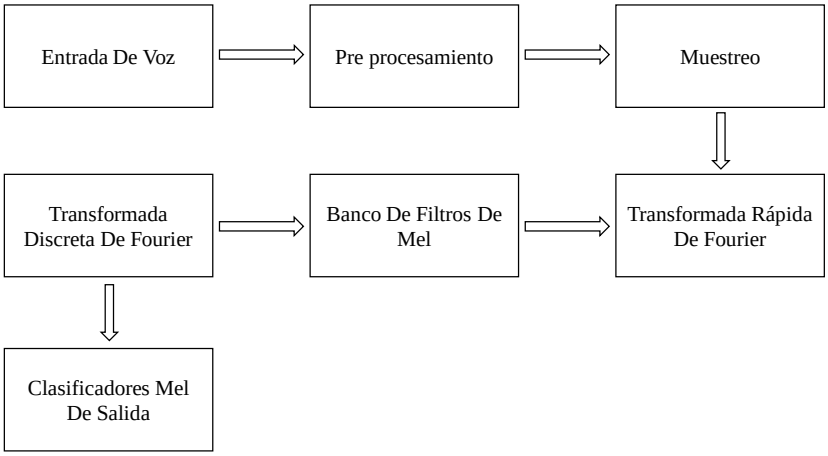

Figura 2.2. Algoritmo de MFCC [19]

Con base en la figura 2.2 y con una red neuronal, en el artículo [18] se realiza un reconocimiento de las palabras “hola”, “encender”, “apagar”, “arriba”, “abajo” y “adiós” obteniendo la tasa porcentual de reconocimiento presentada en la figura 2.3:

	MFCC	MFCC ADAPTATIVO
Hello	96.7%	98%
Turn on	96.2%	96%
Turn off	96.5%	97.2%
Up	96.4%	97.6%
Down	96.3%	98%
Good bye	96%	97.4%

Figura 2.3. Resultados MFCC Y MFCC adaptativo [18] (Palabras originales en inglés)

En cuanto al segundo factor es posible utilizar un sistema de reconocimiento de voz como se muestra en [18], el cual incorpora varias variables o parámetros incluyendo el tono, la frecuencia y la forma de onda. A partir de esto crearon un algoritmo de seguridad el cual permite

representar el sistema de apertura de una puerta prototipo que únicamente se abre cuando la voz del individuo que quiere abrir concuerda con la voz registrada anteriormente.

De igual manera se han realizado distintos métodos para optimizar el reconocimiento de hablantes combinando el uso de los MFCC *Mel Frequency Cepstral Coefficients* junto con diferentes técnicas de aprendizaje automático. El algoritmo que más se encuentra en la literatura de reconocimiento de voz es el de MFCC-GMM el cual usa la extracción de coeficientes de Mel, realiza un modelo de mezcla gaussiana que luego en la fase de prueba compara junto a las características de la entrada. Este algoritmo no requiere de mucha capacidad computacional para funcionar y obtiene una precisión superior al 90 % siempre y cuando se tengan al menos 10 segundos de habla para entrenamiento.[22]

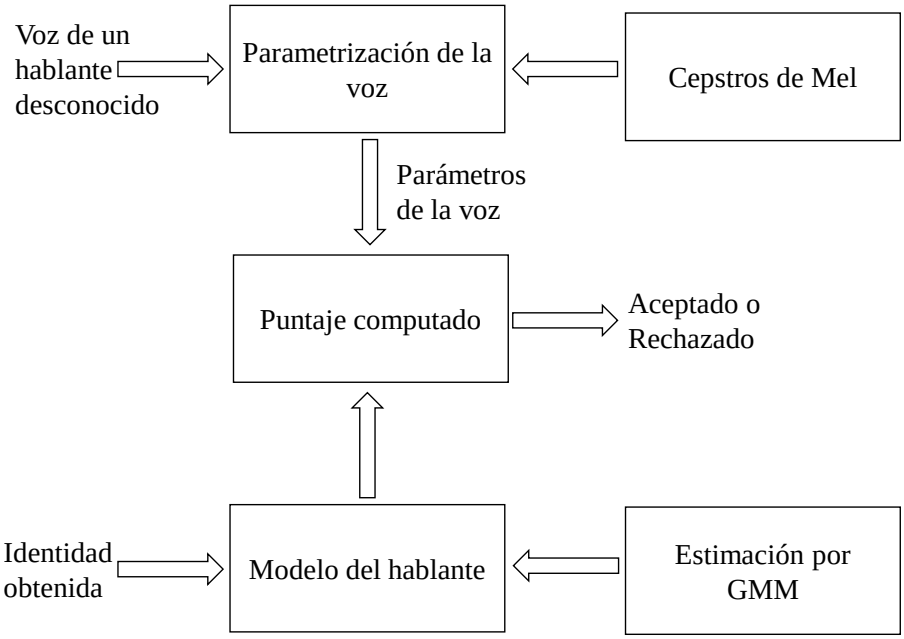

Figura 2.4. Fases de prueba [22]

Con la extracción de los coeficientes de Mel también se han realizado algoritmos que permiten diferenciar si una voz es masculina o femenina.

Para esto es necesaria una base de datos amplia que contenga múltiples audios de las dos clases que se desean entrenar y de igual manera que el algoritmo anterior se realiza la extracción de las características MFCC junto al modelo probabilístico gaussiano que permite identificar subpoblaciones dentro de una gran población. Este sistema tiene una exactitud del 89 % [22].

En los últimos años se han implementado novedosos algoritmos para la verificación del hablante, un ejemplo a ello es el algoritmo FPA *(Flower Pollination Algorithm)* el cual simula el fenómeno natural de la polinización de las flores en primavera [23]. Este algoritmo sigue tres reglas para su implementación: Su búsqueda está basada en los vuelos de Levy [23] , su búsqueda local es simulada como sin vida y autopolinizable y finalmente como en la naturaleza, el polen es transportado por el viento, que es no predecible, en el algoritmo, esto se simula usando factor aleatorio. De esta manera el algoritmo puede encontrar las características más importantes del espectrograma local y globalmente, con esta información se encuentra una trayectoria diferente para cada muestra de entrada, lo que afirman, mejora drásticamente la eficiencia de reconocimiento.

Por otra parte, también existe el reconocimiento de personas por características soft-biométricas como lo son el color de cabello, el color de piel, el color de ojos y la altura. En cuanto al color de piel es posible el reconocimiento utilizando distintos modelos como el Gaussiano que utiliza las funciones de densidad Gaussianas para calcular las probabilidades de que un color de píxel sea un color de piel. Existe otro método el cual extrae un rectángulo de la imagen para identificar si el color puede ser considerado un color de piel y la clasificación de color que utiliza el histograma para separar la piel del resto de la imagen [20]. En la figura 2.5, se muestran los resultados obtenidos en la cual se comparan dichas técnicas:

Figura 2.5. Resultados identificación por color de piel: a)imagen original, b) resultados por modelo de rectángulo, c) resultados por modelo Gaussiano, d) resultados por método de clasificación de color [20]

Como se observa en los artículos anteriormente mencionados, el problema de la re-identificación de personas se intenta solucionar de distintas maneras utilizando un solo método ya sea por identificación facial, identificación por voz o identificación por características soft-biométricas, teniendo buenos resultados pero solucionando el problema parcialmente. Por lo anterior el presente proyecto propone un algoritmo el cual integre los métodos dichos con anterioridad, utilizando algunas de las técnicas y algoritmos expuestos a lo largo del documento con el fin de realizar una re-identificación de personas con mayor eficacia.

Capítulo 3

Justificación

El presente proyecto pretende desarrollar e implementar un algoritmo de re-identificación multimodal, es decir, que integre distintas técnicas utilizadas actualmente, con el fin de lograr una identificación y posterior almacenamiento de parámetros extraídos de imágenes y de sonidos de humanos de forma eficaz, en entornos domésticos. No obstante, que durante los últimos años países tales como Estados Unidos [24], Alemania, Inglaterra [25], China y Japón, entre otros, han realizado desarrollos de robótica social para hogares, en Colombia no es frecuente encontrar esta clase de tecnologías; es por esto que uno de los objetivos de este proyecto es incentivar este tipo de desarrollos, a mediano o largo plazo en el país.

Asimismo, uno de los principales objetivos de la robótica social, consiste en que un robot logre interactuar y comunicarse con seres humanos, siguiendo comportamientos y normas sociales [7]; debido a esto un factor importante en dicha disciplina es la re-identificación de personas. Para lograr lo mencionado, existen diferentes técnicas como la identificación por características biométricas, que son aquellos rasgos fisiológicos que hacen único a un ser humano. como por ejemplo el rostro, la huella dactilar, propiedades de la voz, características de la retina, etc. Pero también existen las características soft-biométricas, que a pesar de no proporcionar alta capacidad para distinción, son de gran ayuda para

definir la identidad de las personas. Algunos ejemplos de dichos rasgos son: el color de la piel, el tinte del cabello, la textura de la ropa, el tamaño, la manera de caminar, posibles cicatrices, tatuajes, etc [26].

Con base a lo descrito en los anteriores párrafos, el presente proyecto propone integrar algunas características biométricas tales como el tono de la voz, la forma del rostro, y algunas otras características soft-biométricas como el color de los ojos, el color del cabello, el color de piel y la estatura. Dado que un elevado porcentaje de trabajos realizados están enfocados en la identificación por un método particular [9][19][20], son escasos los proyectos encontrados en las bases de datos consultadas (IEEE y ScienceDirect), centrados en el mecanismo de la re-identificación multimodal, lo cual podría presentar resultados confiables, respecto de lo realizado en la actualidad. Dicho problema se puede evidenciar en el artículo [9] en donde se menciona únicamente identificación facial por medio de aprendizaje profundo, por lo que se llega a la conclusión de que aunque el algoritmo funciona correctamente, para el caso de personas muy semejantes facialmente, se presenta un elevado porcentaje de error, puesto que el sistema no dispone de parámetros adicionales que le permita lograr diferenciarlas.

De igual manera, se aspira a que con una re-identificación de personas como la anteriormente descrita, el robot sea capaz de cumplir órdenes en una casa familiar de máximo cinco integrantes logrando así que la asistencia doméstica personalizada mejore y por ende la calidad de vida de las personas. Por otro lado, cabe aclarar que el algoritmo que se implementará podrá tener distintas aplicaciones a la mencionada como por ejemplo en vigilancia, rescate, asistencia en eventos sociales, etc. Asimismo, se espera que el presente constituya un aporte importante a la comunidad investigativa, específicamente al Grupo de Estudio y Desarrollo en Robótica - GED de la Universidad Santo Tomás.

Capítulo 4

Objetivos

4.1. Objetivo General

Implementar un algoritmo de re-identificación multimodal, con el fin de que el sistema sea capaz de cumplir órdenes básicas personalizadas a diferentes miembros de una casa familiar.

4.2. Objetivos Específicos

- Recolectar información de diferentes bases de datos acerca de las distintas técnicas y algoritmos utilizados actualmente para la re-identificación de personas.
- Elegir las técnicas y algoritmos que se utilizaran para la realización del proyecto teniendo en cuenta la velocidad y precisión de reconocimiento.
- Implementar un algoritmo que integre las distintas técnicas y algoritmos anteriormente seleccionados.
- Comparar el algoritmo multimodal implementado con los consultados inicialmente que se enfocan en solo una estrategia de identificación, por medio de matrices de confusión y Benchmarks.

- Realizar la validación del sistema por medio de distintas pruebas en las cuales se evaluará que el sistema cumple órdenes eficientemente dependiendo de la persona en un ambiente familiar

Capítulo 5

Marco Teórico

Para el presente proyecto es necesario contextualizar algunos conceptos como los que se pueden apreciar a continuación:

Histograma de Gradientes:*(Histogram of Oriented Gradients HOG)* este algoritmo es utilizado para localizar rostros dentro de una imagen, el procedimiento que utiliza es el siguiente: en primer lugar la imagen es convertida a blanco y negro, luego se recorre cada pixel de la imagen observando los pixeles que están alrededor y se dibuja una flecha que muestra la dirección en que oscurece la imagen [27].

Si se repite ese proceso para cada píxel en la imagen, terminará reemplazando cada píxel por una flecha. Estas flechas se llaman gradientes y muestran el flujo de claro a oscuro en toda la imagen. Como guardar la dirección de cada píxel da mucho detalle, lo que se hace es agrupar en cuadros de 16x16 píxeles y se toma la dirección del gradiente que más se repita. En la figura 5.1 se presenta la imagen resultante del proceso:

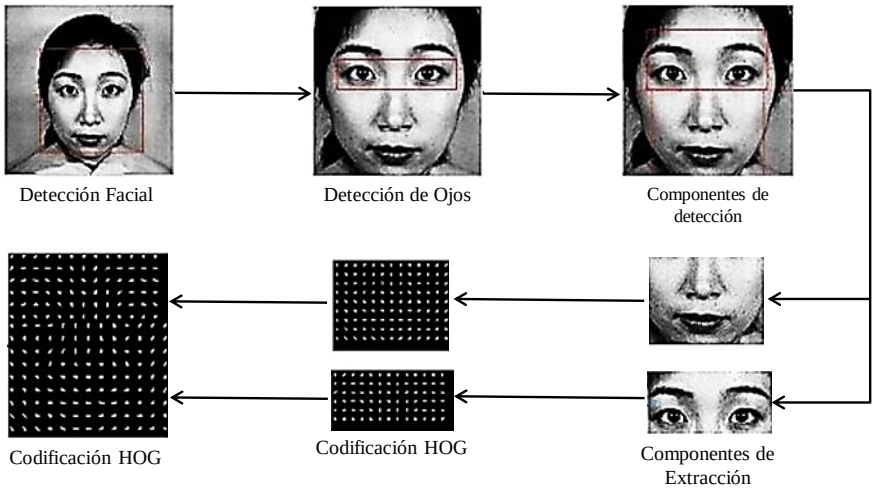

Figura 5.1. Procedimiento algoritmo Hog [27]

Finalmente, selecciona la parte de la imagen con más similitud a un HOG previamente entrenado

Viola Jones: Fue el primer algoritmo de detección de objetos en tiempo real, sin embargo actualmente se utiliza en su mayoría para detección facial, fue propuesto en el año 2001 por Paul Viola y Michael Jones. El procedimiento que emplea es el siguiente: en primer lugar convierte la imagen a escala de grises, posteriormente se divide la imagen en subregiones que se comparan con una serie de características preestablecidas llamadas características Haar,en la figura 5.2 se presentan algunos ejemplos de dichas características:

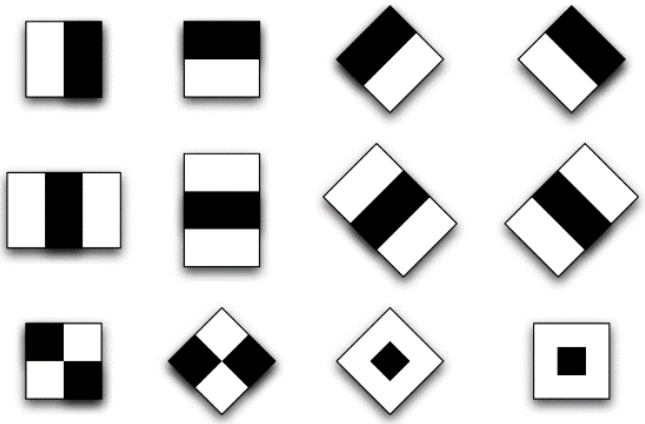

Figura 5.2. Clasificadores Haar[28]

Si la subregión no es similar al clasificador, automáticamente ese pedazo de la imagen no se clasifica como rostro y se continúa con el resto, de esa manera se ahorra tiempo y recursos ya que solo se procesa las subregiones en las que se cree que hay un rostro [28].

Máquinas de soporte vectorial: *(Support Vector Machines SVM)* reúnen una serie de algoritmos desarrollados en la década de los 90 por Vladimir Vapnik. Su funcionamiento se basa en el concepto de hiperplano, el cual busca realizar una separación entre las diferentes clases, los SVM´s buscan maximizar el margen entre las clases, es decir, la frontera de decisión debe estar a la mayor distancia posible de las clases que esté separando.
La ecuación 5.1 hace referencia a un hiperplano:

$$(\beta_0 + \beta 1_{x1} + \beta 2_{x2} + ... + \beta p_{xp} = 0) \tag{5.1}$$

El valor del hiperplano puede tomar valores positivos o negativos, haciendo referencia esto a cada una de las clases. Es aquí donde surge el concepto de confiabilidad, entre más positivo sea el valor tendrá más confiabilidad de ser de una clase y entre más negativo, más confiabilidad habrá de que sea de la otra.

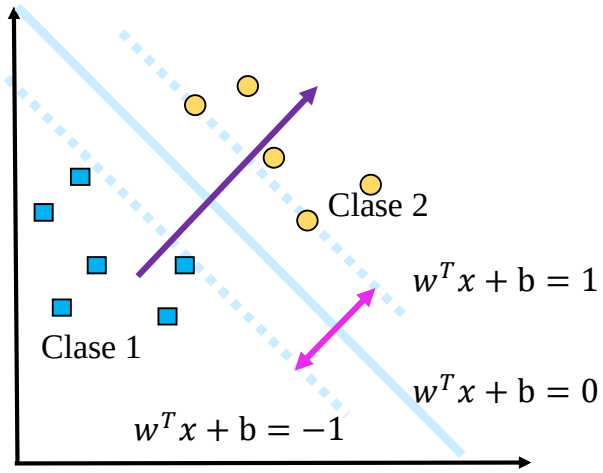

Figura 5.3. Margen de separación Support Vector Machine [29]

Sin embargo, muchas veces la separación entre las clases no se puede realizar de manera lineal, para estos casos es necesario hacer curvas no lineales de separación así como agregar *Kernel* que permite hacer más flexible el hiperplano construido [29].

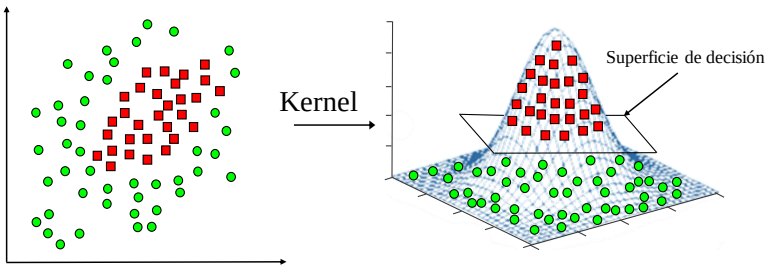

Figura 5.4. Kernel Support Vector Machine [29]

Radial SVM:El comportamiento del kernel se explica por medio de la ecuación 5.2:

$$K(x, y) = (exp(-\gamma \sum_{j=1}^{p}(x_{ij} - y_{ij})) \tag{5.2}$$

Si γ es muy grande, entonces se obtienen límites de decisión fluctuantes y ondulantes silenciosos, lo que explica la alta variación y el sobreajuste. Si γ es pequeño, la línea de decisión o el límite es más suave y tiene poca varianza.
Entonces, ahora la ecuación del clasificador de vectores de soporte se convierte en:

$$F(x) = (\beta_0 + \sum_{y_0 S}(\alpha y_0 K(x y_0, y y_0)) \tag{5.3}$$

donde S son los vectores de soporte y α es un valor de peso distinto de cero para todos los vectores de soporte [30].

Redes neuronales: Las redes neuronales son modelos aproximados del funcionamiento del sistema nervioso. Las unidades básicas son las neuronas, que generalmente se organizan en capas [31], como se muestra en la figura 5.5.

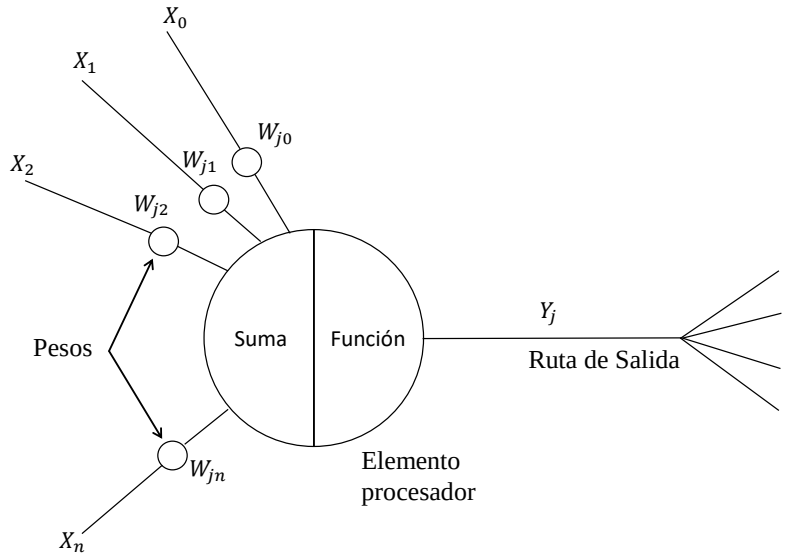

Figura 5.5. Modelo de redes neuronales [31]

Como se puede observar en la imagen 5.5, un modelo de redes neuronales se conforma principalmente por tres partes: la capa de entrada en la cual se reciben los datos reales que alimentan a la red neuronal,las capas ocultas en las cuales se desconocen los valores tanto de entrada como de salida y finalmente la capa de salida donde se muestra el resultado de la red. El entrenamiento de una red neuronal, se realiza ajustando los pesos de las entradas, de tal forma que la salida sea lo más parecida a los datos o valores deseados [31].

Aprendizaje profundo: Es una rama del aprendizaje automático que utiliza redes neuronales que funcionan muy parecido a las conexiones neuronales del cerebro humano.

El término "profundo" suele hacer referencia al número de capas ocultas en la red neuronal. Las redes neuronales tradicionales sólo contienen dos o tres capas ocultas, mientras que las redes profundas pueden tener una

gran cantidad de dichas capas ocultas [1].

K vecinos más cercanos: (k-nearest neighbors knn) es un algoritmo de clasificación supervisada, este método busca las observaciones más cercanas a la clase que está intentando predecir y clasifica el punto de interés a partir de los datos que lo rodean . Su funcionamiento se puede reducir en los siguientes pasos: en primer lugar calcula la distancia entre el dato de prueba y los entrenados, luego selecciona los "k" elementos más cercanos y finalmente decide a qué categoría pertenece dependiendo de la predicción que más se repita en el grupo de vecinos [32].

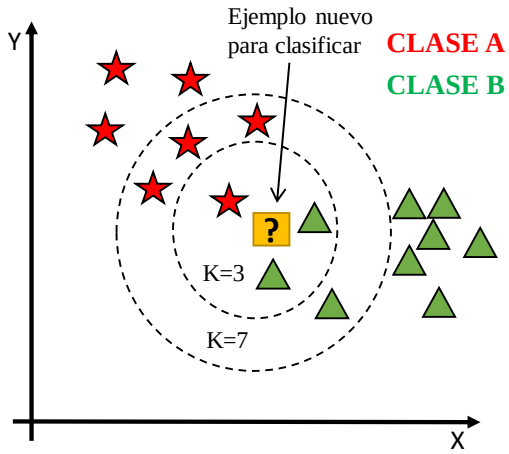

Figura 5.6. Algoritmo Knn [32]

Framing: La señal de la voz es no estacionaria, ya que su frecuencia está continuamente cambiando en el tiempo, para realizar un análisis de la voz es necesario recortar el contenido en pequeños intervalos de tiempo y de esta manera poder ver la señal como lineal e invariante en el tiempo [33].

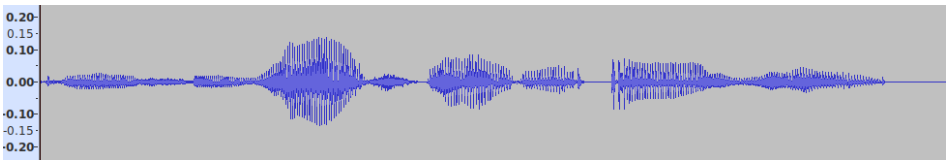

Figura 5.7. Señal de voz, No estacionaria(fuente autor)

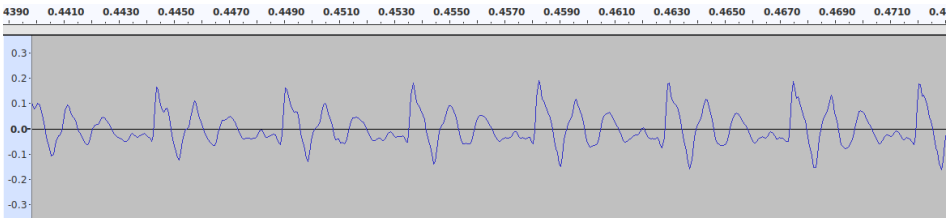

Figura 5.8. Señal de voz, No estacionaria en un intervalo de 25ms(fuente autor)

Windowing: La extracción de cuadros sin procesar de una señal de voz puede provocar discontinuidades hacia los puntos finales debido a un número no entero de períodos en la forma de onda extraída, lo que conducirá a una representación de frecuencia errónea (conocida como fuga espectral en procesamiento de señales). Esto se evita multiplicando una función de ventana con la trama de la voz. La amplitud de una función de ventana cae gradualmente a cero hacia sus dos extremos y, por lo tanto, esta multiplicación minimiza la amplitud de las discontinuidades mencionadas anteriormente [33].

Superposición de fotogramas: Debido a la creación de ventanas, en realidad se están perdiendo las muestras hacia el principio y el final del fotograma; esto también conducirá a una representación de frecuencia incorrecta. Para compensar esta pérdida, se toman fotogramas superpuestos en lugar de fotogramas disjuntos, de modo que las muestras perdidas desde el final del i-ésimo fotograma y el comienzo del (i + 1)-ésimo fotograma se incluyen por completo en el fotograma formado por la superposición entre estos 2 marcos. La superposición entre tramas se considera generalmente de 10 a 15 ms.

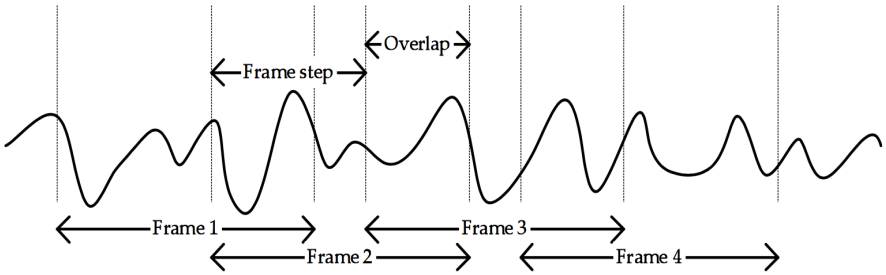

Figura 5.9. Superposición de frames en una señal de voz [33]

Coeficientes Cepstrales de las frecuencias de Mel: *(Mel Frequency Cepstral Coefficients)* son coeficientes para la representación del habla basados en la percepción auditiva humana. Los MFCC muestran las características locales de la señal de voz asociadas al tracto vocal y el *pitch* o vibración de las cuerdas vocales [19]. El habla es producida por los seres humanos mediante un filtro aplicado por nuestro tracto vocal sobre el aire expulsado por los pulmones. Las propiedades de la fuente (pulmones) son comunes para todos los seres humanos; son las propiedades del tracto vocal las que se encargan de dar forma al espectro de la señal y varía según la persona. La forma del tracto vocal define el sonido que se produce y los MFCC representan mejor estas ondas de audio. Para poder realizar el procesamiento se necesitaría que el sistema sea LTI (lineal e invariante en el tiempo); como la voz no corresponde a un sistema de este tipo, es necesario tomar pequeños fragmentos de audio aproximadamente de 20-25 ms donde se asemeje a un sistema de este tipo [33].

A partir de la teoría de la producción del habla, se supone que el habla es la convolución de la fuente (aire expulsado de los pulmones) y el filtro (nuestro tracto vocal). El propósito aquí es caracterizar el filtro y eliminar la fuente.

Para lograr esto, primero se transforma la señal de voz en el dominio del tiempo en una señal en el dominio espectral usando la transformada de Fourier, donde la fuente y la parte del filtro están ahora en multiplicación, seguido de esto se toma el registro de los valores transformados para que la fuente y el filtro sean ahora aditivos en el dominio espectral. El uso de logaritmo para transformar de multiplicación a suma facilita la separación de la fuente y el filtro mediante un filtro lineal.
Finalmente, se aplica la transformada de coseno discreta (que resulta ser más exitosa que FFT o I-FFT) de la señal espectral logarítmica para obtener MFCC. Inicialmente, la idea era transformar la señal espectral logarítmica en el dominio del tiempo usando FFT inversa, pero el logaritmo, al ser una operación no lineal, crea nuevas frecuencias llamadas Quefrency.

La razón del término "mel" en MFFC es la escala mel, que especifica exactamente cómo espaciar nuestras regiones de frecuencia. Los seres humanos son mucho mejores para discernir pequeños cambios de tono a bajas frecuencias que a altas frecuencias. La incorporación de esta escala hace que nuestras características coincidan más con lo que escuchan los humanos [35].

Escala de Mel: La escala de mel es el resultado de la relación entre la frecuencia real y la frecuencia percibida por los seres humanos. Esta fue hecha por Stevens y Volkman. Esta escala ha sido ajustada con el paso de los años por otras personas como Koening, Fant y O'Shaughnessy, para este trabajo fue usada la fórmula propuesta por O'Shaughnessy; puesto que es la más aceptada en la comunidad científica [34].

$$m = (2595(\log_{10}(1 + \frac{f}{700})) = 1127 \ln(1 + \frac{f}{700})) \tag{5.4}$$

Modelo de Mezcla Gaussiana: Un modelo de mezcla gaussiana *GMM* es una función de densidad de probabilidad paramétrica representada como una suma ponderada de las densidades de los componentes

gaussianos. Los GMM se utilizan comúnmente como un modelo paramétrico de la distribución de probabilidad de medidas o características continuas en un sistema biométrico, como las características espectrales relacionadas con el tracto vocal en un sistema de reconocimiento de hablante [36].

Segmentación de color por histograma y agrupamiento k-medias: Para el reconocimiento de color de piel, se utilizó el método utilizado en el articulo [37], en el cual se realiza el siguiente procedimiento:
- Segmentación de la imagen: Hay varios métodos para realizar la segmentación de imágenes, incluidos los métodos de umbralización, agrupación, transformación y textura, en este caso se emplea la segmentación por umbral basado en histogramas, para esto en primer lugar la imagen rgb es convertida a escala de grises, de esta manera los píxeles se dividen en función de sus valores de intensidad y los resultados deben revelar valores que separan el fondo de la imagen del objeto de interés, que en este caso es la persona que se desea reconocer.

- Detección de piel: Debido a que el formato RGB se considera sensible a los cambios de luz, con el fin de lograr mejores resultados se utilizan los espacios de luz HSV (Hue, Saturation, Value) donde el valor de H define el tono del color y los valores de S y V la saturación y el brillo respectivamente, de igual manera se utiliza YCrCb donde el valor de Y determina la luminosidad de la imagen y las señales CB y CR son los componentes de crominancia diferencia de azul y diferencia de rojo.

- En este paso se utiliza la ecuación 5.5 para determinar si un píxel:

$$piel = (1, si\ CR \geq 140\ y\ CR \leq 140\ y\ CR \geq 140\ y\ CR \leq 170\ y\ H < 0{,}95) \tag{5.5}$$

- Por último se utiliza K-Means Clustering que es un método simple que agrupa los píxeles en grupos definidos a partir de la medida euclidiana

cuadrada como distancia, en este caso se tienen los siguientes grupos: píxeles de fondo, primer plano o persona y piel

Red neuronal de convolución basada en regiones *(Region Based Convolutional Neural Networks RCNN)*: estas redes se utilizan bastante en la detección de objetos, para crear un límite en cada objeto presente en la imagen y básicamente consta de dos pasos: en primer lugar consiste en la extracción de vectores de características y un conjunto de SVM lineales, con el fin de no tener que clasificar cada uno de los píxeles de la imagen, se utiliza una búsqueda selectiva para solo extraer 2000 regiones, luego con las regiones obtenidas se alimenta la red convolucional que extrae las características de cada región y utiliza SVM para generar las etiquetas de cada clase [38].

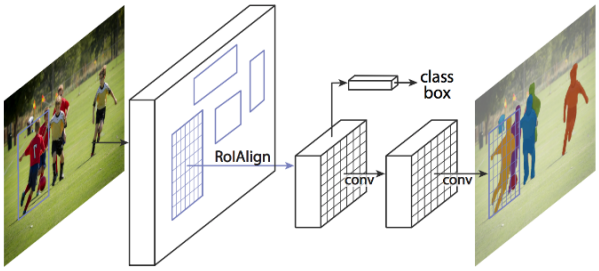

Figura 5.10. Red neuronal convolucional basada en regiones [38]

Red convolucional en cascada multitarea MTCNN: Es un frame desarrollado en python que se basa en el artículo [39] que busca solucionar el problema tanto de detección como de alineación de rostros. El proceso consta de tres etapas de redes convolucionales que logran reconocer tanto rostros como ojos, nariz y boca.

El documento propone MTCNN como una forma de integrar ambas tareas (reconocimiento y alineación) utilizando el aprendizaje multitarea, En la primera etapa, utiliza una CNN poco profunda para producir rápidamente ventanas candidatas. En la segunda etapa refina las ventanas

de candidatos propuestos a través de una CNN más compleja. Y por último, en la tercera etapa utiliza una tercera CNN, más compleja que las demás, para refinar aún más el resultado y generar posiciones de referencia facial.

Teorema de Bayes: Es una ecuación que describe la relación de probabilidades condicionales, dicho en otras palabra indica la probabilidad de que ocurra un evento dadas unas ciertas condiciones o características observadas [40], ejemplo: cuál es la probabilidad de que llueva dado que el cielo está nublado, lo anterior se puede descubrir por medio de la ecuación 5.6:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{5.6}$$

Donde:
- P(A): probabilidad a priori
- P(B—A): es la probabilidad de B en la hipótesis A
- P(A—B): es la probabilidad a posteriori (y resultado a encontrar)

Función Softmax: también conocida como función exponencial normalizada, se encarga de asignar probabilidades decimales a cada clase en un caso de clases múltiples [41], de tal manera que la suma de dichas probabilidades sea igual a 1, puede ser utilizada para representar una distribución categórica– la distribución de probabilidad sobre K diferentes posibles salidas. La función está dada por:

$$\sigma(z)j = \frac{\exp^{zj}}{\sum_{k=1}^{K} \exp^{zk}} \tag{5.7}$$

Dónde:
-K es el tamaño del vector original
-zj los valores de cada posición del vector
-σ(z)j es el vector resultante con valores en un rango de 0 a 1

Reconocimiento del habla *(speech to text)*: Es una rama de la Inteligencia Artificial que se encarga de permitir la comunicación hablada entre los seres humanos y las computadoras. Este sistema de reconocimiento de voz es capaz de procesar una señal de audio emitida por una persona y convertirla en texto para su posterior análisis [42].

Sistema Operativo Robótico *(Robot Operating System ROS)*: Es una colección de herramientas, bibliotecas y convenciones que tienen como objetivo simplificar la tarea de crear un comportamiento robótico complejo y robusto en una amplia variedad de plataformas robóticas [43].

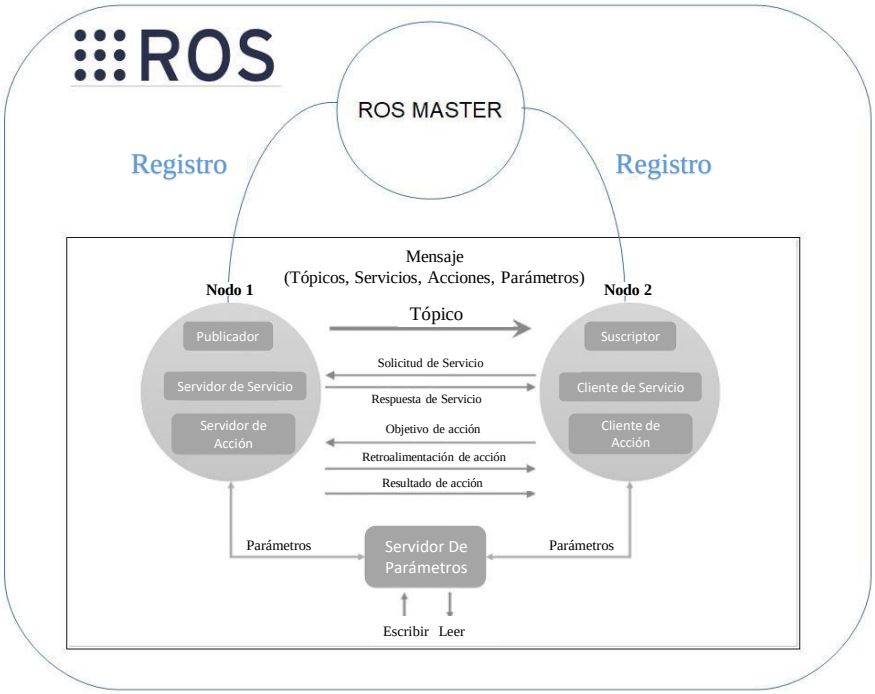

Figura 5.11. Estructura de ROS[43]

A continuación se explican algunos términos clave utilizados en el desarrollo del proyecto con el sistema operativo robótico:

Paquete: Los paquetes en ROS son los elementos principales de su arquitectura, debido que proporcionan una capa de comunicación para facilitar la transmisión entre los nodos, siendo estos los paquetes que pueden contener los procesos de ROS en tiempo de ejecución, conjunto de datos y archivos de configuración [46].

Nodo: es un proceso que realiza cálculos, en otras palabras es un archivo que se puede comunicar con otros archivos. El uso de nodos en ROS proporciona varios beneficios al sistema en general, como por ejemplo existe una tolerancia a fallos adicional ya que los fallos se aíslan en nodos individuales. Asimismo, la complejidad del código se reduce en comparación con los sistemas monolíticos [43].

Mensaje: Es la forma de comunicación entre nodos en ROS y básicamente, es una estructura de datos simple, que comprende campos escritos. Se admiten los tipos primitivos estándar (entero, punto flotante, booleano, etc.), al igual que las matrices de tipos primitivos [43].

Servicio: es un paradigma de comunicación muy flexible, en el cual un nodo ROS proveedor ofrece un servicio con un nombre de cadena , y un cliente llama al servicio enviando el mensaje de solicitud y esperando la respuesta[43].

Capítulo 6

Desarrollo Metodológico

Lo primero que se realizó fue una consulta bibliográfica de los diferentes algoritmos para cada uno de los métodos de reconocimiento de personas, en las tablas 6.1, 6.2, 6.3 y 6.4 se muestran los resultados:

Detección facial			
Algoritmo	**Porcentaje**	**Ventajas**	**Desventajas**
Viola Jones[44]	71.11 %	-Puede detectar muy bien la cara frontal en las imágenes.	Bastante difícil de detectar la cara cuando la persona usa casco, gafas o mascaras.
Hog[44]	79 %	-Detecta mejor que VJ caras en diferentes posiciones y escalas	Bastante difícil de detectar la cara cuando la persona usa casco, gafas o mascaras
OpenFace (Hog y Svm)[45]	92.92 %	Puede manejar una mala iluminación y varias posiciones faciales.	

Tabla 6.1. Métricas de evaluación adultos detección facial

Reconocimiento facial					
Algoritmo	**Acc**	**Pre**	**Rec**	**Ventajas**	**Desventajas**
KNN[46]	66 %	83 %	67 %	Simplicidad del sistema	Dependiendo la cantidad de clases puede representar gran coste computacional
SVM[46]	83 %	83 %	83 %	Mejor comportamiento en cuanto accuracy, precision, recall frente a KNN	Puede ser ineficiente cuando se tiene gran cantidad de clases a entrenar

Tabla 6.2. Métricas de evaluación adultos detección facial

Reconocimiento por Voz			
Algoritmo	**Porcentaje**	**Ventajas**	**Desventajas**
MFCC-GMM [47][48][49]	95 %-96.8 % [47][52]	-Reconocimiento sin importar texto de entrada eficiente computacionalmente. Mejor aproximación para densidades de formas arbitrarias. [48] - Facilidad de implementación -Se puede entrenar con datos de entrada limitados menor a 15s [49].	- Su porcentaje de acierto baja cuando la voz del hablante es modificada por sus emociones [48].
MFCC-SVM [48]	92.5 % [48]	- Facilidad de implementación	- Falla cuando los audios tienen ruido [51] - Su porcentaje de acierto baja cuando la voz del hablante es modificada por sus emociones [48] -El porcentaje de acierto es baja cuando se aumenta el número de hablantes del sistema. [54] -Usando principalmente como clasificador binario[55].

MFCC-MLP [48]	92 % [48]	- No se evidencian ventajas sobre los demás algoritmos.	- Se evidencia dificultad para encontrar una implementación accesible. - Su porcentaje de acierto baja cuando la voz del hablante es modificada por sus emociones [48].
MFCC-ResNet [50]	88 % [50]	- Capacidad para ser usado con cientos de hablantes diferentes, por lo que lo hace óptimo para un sistema complejo [53]. - Facilidad de implementación.	- Costo computacional (Uso de GPU).

Tabla 6.3. Consulta bibliográfica algoritmos Reconocimiento de voz

Reconocimiento softbiométrico			
Algoritmo	**Porcentaje**	**ventajas**	**Desventajas**
color segmentation with histogramand K-means clustering[37]	68,1 %	El metodo detecta el color de la piel con una precisión razonable y tiene bajo costo compuatcional	Falla en imagenes donde la piel ocupa pocos pixeles.
R-CNN[56]	69,1 %	Resultados de detección razonables	Tiene un costo computacional alto, logra rendimiento decente utilizando GPU.
Red convolucional en cascada multitarea MTCNN.[56]	69 %	Resultados de detección razonables	Tiene un costo computacional alto, logra rendimiento decente utilizando GPU.

Tabla 6.4. Métricas de evaluación adultos detección facial

A continuación se presenta el detalle de cada una de las etapas del proyecto, asimismo se explica cada una de las pruebas realizadas para elegir los diferentes algoritmos a utilizar con el fin de implementar un algoritmo multimodal de re-identificación de personas, finalmente los resultados de cada una de las pruebas planteadas, se pueden observar en el capitulo 7 de resultados.

Para comenzar, es necesario tener presente que el proyecto se divide en tres partes fundamentales: la primera es la selección de algoritmos en cada

una de las estrategias de re-identificación: Reconocimiento facial, por voz y por características soft-biométricas, la segunda es la integración de los algoritmos anteriormente seleccionados en un único desarrollo y finalmente la implementación del algoritmo desarrollado en el paso anterior, en un entorno doméstico. En la figura 6.1, se puede observar el procedimiento utilizado en el presente proyecto.

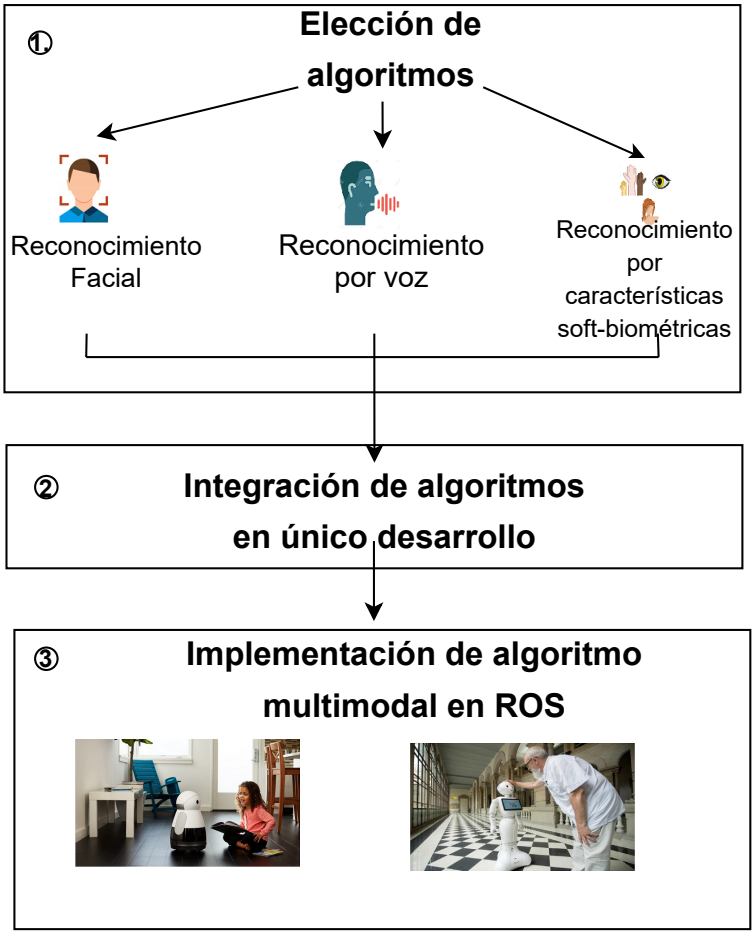

Figura 6.1. Metodología proyecto

6.1. Reconocimiento facial

6.1.1. Detección facial

Antes de llevar a cabo el proceso de re-identificación facial, se debe localizar el rostro o los rostros que hay dentro de la imagen, es por esto que en cuanto a detección facial se probaron y evaluaron tres métodos: HOG, Viola Jones y el algoritmo OpenFace que utiliza una combinación entre HOG y SVM.

Cada uno de los algoritmos se evaluaron a partir de matrices de confusión, tal como se muestra en la tabla 6.5:

	Predicción	
Observación	Positivos	Negativos
Positivos	VP	FN
Negativos	FP	VN

Tabla 6.5. Matriz de confusión

Los parámetros mostrados en la tabla 6.5 se pueden definir de la siguiente manera:

VP *(Verdaderos Positivos)*: indica el número de caras que el algoritmo detectó correctamente.
VN *(Verdaderos Negativos)*: indica el número de regiones que el algoritmo no detectó como caras correctamente.
FP *(Falsos Positivos)*: indica el número de regiones que el algoritmo detectó como cara incorrectamente
FN *(Falsos Negativos)*: Indica el número de caras que el algoritmo no detectó.

En la figura 6.2 se presenta un pequeño resumen del procedimiento para evaluar los algoritmos de detección facial, asimismo más adelante se explica a detalle cada una de las fases.

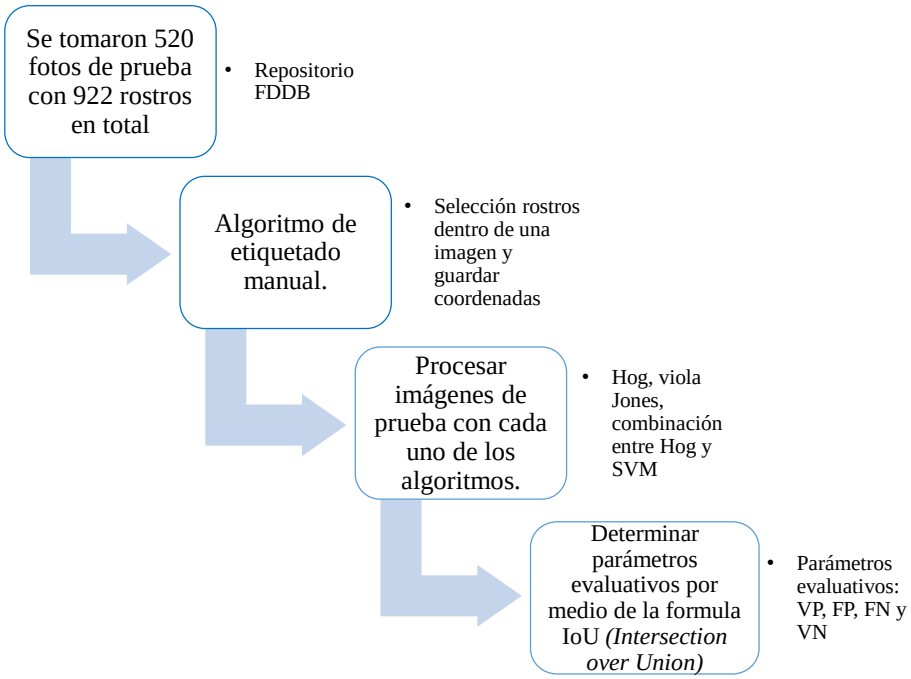

Figura 6.2. Procedimiento detección facial.

De este modo, lo primero que se realizó fue un algoritmo el cual permitía seleccionar los positivos de observación, es decir los rostros que realmente había en una imagen y guardar sus correspondientes coordenadas.

Figura 6.3. Detección facial[48]

Para las pruebas, se utilizaron 520 fotos obtenidas del repositorio FDDB (Face Detection Data Set and Benchmark) [48] en las cuales aparecían 922 rostros en total, estos fueron etiquetados a mano para su posterior análisis. Luego a cada uno de los algoritmos se les pasó las imágenes de prueba y de igual manera se guardaron las coordenadas de la detección, con el fin de obtener tanto los rostros reales como los rostros predichos, de la siguiente manera:

En la figura 6.4 se puede observar el rostro seleccionado a mano de color verde, asi como el rostro predicho por el algoritmo de color azul. De esta manera se calculó la superposición de los cuadros, con el fin de determinar si se trataba de un verdadero positivo, un falso positivo o un falso negativo, por medio de la fórmula Intersection over Union (IoU) tal como se muestra en la ecuación 6.1:

$$IOU = \frac{Area\ of\ overlap}{Area\ of\ union} \tag{6.1}$$

la formula 6.1, se explica por medio de la figura 6.5:

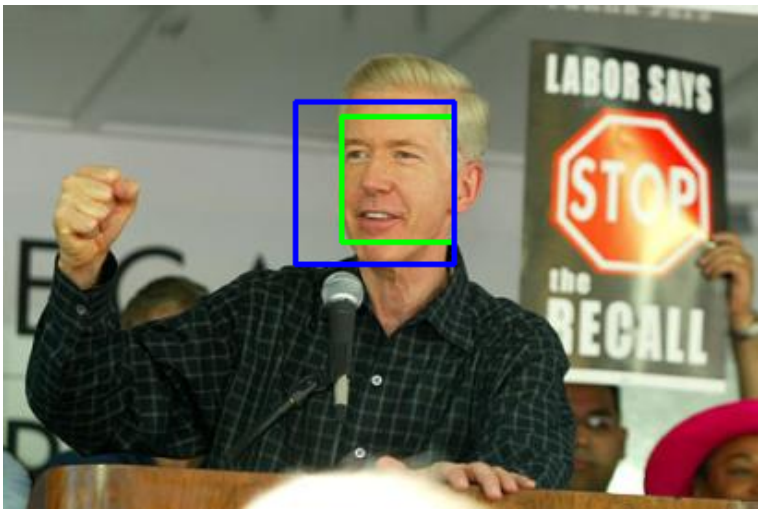

Figura 6.4. Detección real vs predicción[48]

$$IOU = \frac{Area\ de\ solapamiento}{Area\ de\ unión} =$$

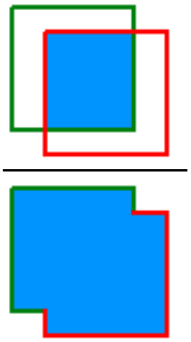

Figura 6.5. Explicación IoU

Con base en la figura 6.5, se evaluó cada uno de los rostros con un umbral de 0.2 definido durante el desarrollo de las pruebas, de esta manera cuando IoU era mayor al umbral se trataba de un verdadero positivo.En la figrua 6.6 se muestra un ejemplo de cómo se obtuvieron los parámetros VP, FP y FN. Por otro lado, en cuanto a los VN, se utilizaron imágenes en las cuales no había ningún rostro, con el fin de evaluar el desempeño de cada algoritmo:

Figura 6.6. Verdadero positivo (Izq), falso positivo (Medio), falso negativo (Der).

Una vez obtenidos los parámetros de la tabla 6.5, se procedió a calcular las métricas de evaluación, teniendo en cuenta las siguientes fórmulas y definiciones:

Precision: Es la proporción de casos predichos positivos que fueron correctos

$$Precision = \frac{VP}{VP + FP} \tag{6.2}$$

Recall: Muestra la cantidad de verdaderos positivos que el modelo ha clasificado en función del número total de valores positivos

$$Recall = \frac{VP}{VP + FN} \tag{6.3}$$

Accuracy: Indica el número de elementos clasificados correctamente en comparación con el número total de artículos, para esta métrica las clases deben estar equilibradas en cantidad de elementos

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} \tag{6.4}$$

6.1.2. Reconocimiento facial

Al igual que con la detección facial, las pruebas de los algoritmos se realizaron con matrices de confusión, a continuación se explican los parámetros evaluados:

Verdadero Positivo (VP): El algoritmo reconoce correctamente la cara test con una de las personas entrenadas anteriormente.

Falso Negativo (FN): El algoritmo reconoce incorrectamente la cara test como desconocida, cuando en realidad si ha sido entrenada.

Falsos Positivos (FP): El algoritmo reconoce incorrectamente la cara test con una de las personas entrenadas anteriormente, es decir confunde a la persona.

Verdaderos Negativos (VN): El algoritmo reconoce correctamente la cara test como desconocida.

Primera prueba:

En la figura 6.7 se presenta un pequeño resumen del procedimiento para evaluar los algoritmos de reconocimiento facial, asimismo más adelante se explica a detalle cada una de las fases.

Se entrenaron en total seis personas, cuatro famosos y los autores del proyecto, asimismo, se utilizaron entre 250 y 300 imágenes para entrenar a cada persona y 100 imágenes de prueba en las cuales aparecen 244 rostros tanto de personas entrenadas como desconocidas. Para el entrenamiento con la librería OpenFace, se realizaron los siguientes pasos, que se pueden encontrar más a detalle en el repositorio https://github.com/TadasBaltrusaitis/OpenFace:

En primera medida, se debía crear un directorio que a su vez contenía subdirectorios con las imágenes de cada persona, de igual manera el nombre de la carpeta debía ser la etiqueta con la cual se iba a identificar al sujeto, cada directorio podía tener diferente cantidad de imágenes. Sin embargo, como se mencionó anteriormente todas las personas fueron entrenadas entre 250 y 300 fotos que podían estar en formato jpg o png.

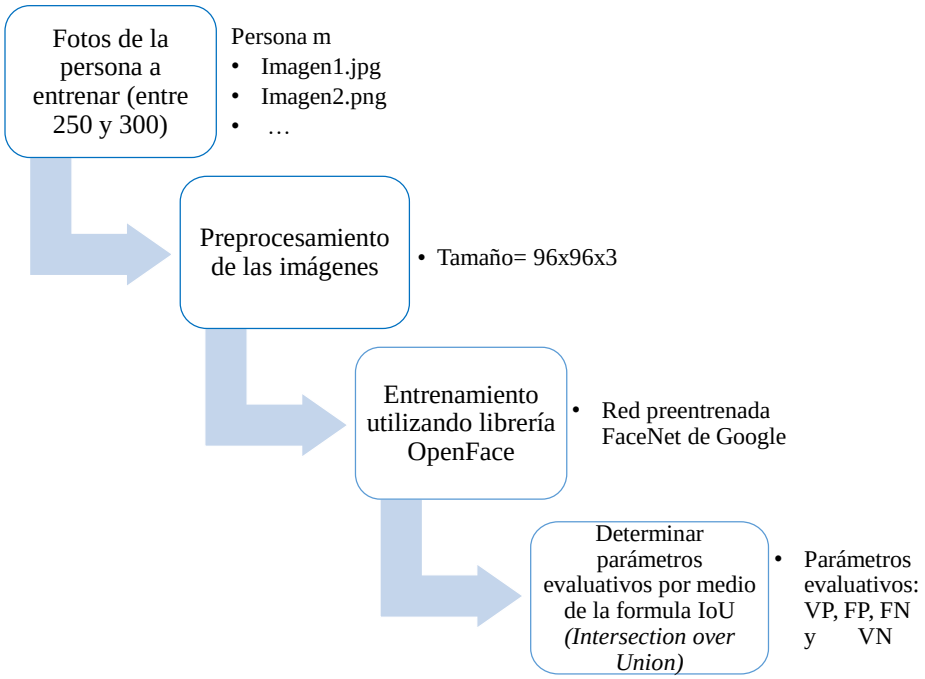

Figura 6.7. Procedimiento Reconocimiento facial

A continuación se presenta la estructura utilizada:

- Persona-1
  - imagen-1.jpg
  - imagen-2.png
  - ...
- Personam
  - imagen-1.jpg
  - imagen-2.png
  - ...

El siguiente paso consistía en detectar el rostro con el algoritmo de mejor desempeño, seleccionado a partir de las pruebas realizadas en el inciso anterior y de los resultados que se muestran en la sección 7.1.1. Del mismo modo se realizaba un pre-procesamiento de las imágenes de tal manera que tuvieran un tamaño de 96x96x3.

Posteriormente se realiza el entrenamiento utilizando la librería de OpenFace con los siguientes parámetros:

output_dim = n (cantidad de imágenes con la que se entrenó a la persona)
input_dim = 128
tamaño_lote = 128
nb_epoch = 100

Sin embargo, cabe aclarar que la librería OpenFace está previamente entrenada y utiliza la arquitectura FaceNet de Google que utiliza 500.000 imágenes para la extracción de características.

De igual manera, el procedimiento para probar los algoritmos, se explica a continuación:

En primer lugar se elaboró un algoritmo el cual permitía etiquetar manualmente las personas que se encontraban dentro de la imagen, de esta manera se guardaban tanto la posición de la cara como el nombre de la persona:

De esta manera, una vez obtenidas las coordenadas y la correspondiente etiqueta de la persona que aparecía en las imágenes de prueba, se procedió a probar cada uno de los algoritmos, cabe aclarar que las posibles clases a reconocer eran las siguientes:

- Lina_Plazas

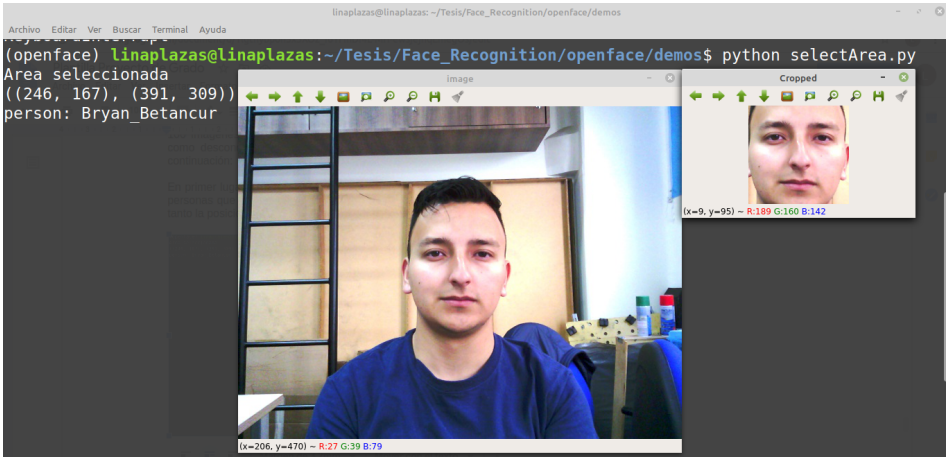

Figura 6.8. Etiquetado manual de imágenes.

- Bryan_Betancur
- Abigail_Breslin
- Adam_Sandler
- Andy_Samberg
- Bella_Thorne
- Unknown

Segunda prueba:

Se realiza el mismo procedimiento nombrado anteriormente, con la diferencia de que dado el propósito del presente proyecto, solamente se evalúan imágenes que contengan una sola persona, de este modo el entrenamiento se realizó con 250 fotos por sujeto y las pruebas con 50 imágenes por individuo, diferentes a las utilizadas en el entrenamiento, adicionalmente se probó con 50 imagenes de desconocidos. Asimismo, se cambiaron las clases a reconocer, con el fin de tener mayor variedad sociocultural, a continuación los nombres de las etiquetas utilizadas:

- Lina_Plazas
- Bryan_Betancur
- Abigail_Breslin
- Adam_Sandler
- Andy_Samberg
- Oprah_Winfrey

6.2. Reconocimiento por voz

Para el reconocimiento por voz primero se implementó un algoritmo de reconocimiento de género, esto con la finalidad de agilizar el proceso de identificación, reduciendo la cantidad de modelos a evaluar a la hora de comprobar a qué persona corresponde la señal de audio de entrada o reconocimiento facial. En segunda instancia se implementó el algoritmo de MFCC-GMM y MFCC-ResCNN respectivamente.

Los algoritmos fueron probados con 60 hablantes diferentes LibriSpeech ASR corpus dev-clean, dev-other, test-clean, test-other para el entrenamiento del reconocimiento de género. De igual manera para la identificaciòn del hablante se usaron 20 hablantes masculinos, 20 femeninos del dataset ya mencionado anteriormente, adicionalmente se usó el dataset Speech Databases of Typical Children[48] que contiene la voz de 20 hablantes menores de edad. Para un total de 60 voces diferentes. Para el algoritmo final se entrenaron 5 personas, 3 mujeres y 2 hombres asumiendo que es un posible caso en una ambiente familiar común. Cada individuo con 100 diferentes audios de 12 a 18 segundos cada audio.

Aunque los dos algoritmos tuvieron resultados similares se decidió escoger el que usa los modelos de mezcla gaussiana puesto que no necesita el uso

de una GPU (Unidad de procesamiento gráfico por sus siglas en inglés) para funcionar rápidamente y el entrenamiento de un nuevo hablante es más sencillo y veloz.

6.3. Reconocimiento por características soft biométricas

Para los algoritmos de reconocimiento de color de piel, color de ojos y color de cabello se realizaron las siguientes pruebas:

Primera prueba:

En la prueba número uno, se seleccionaron minuciosamente fotografías de las personas, las cuales tuvieran en primer lugar buena resolución, además de estar totalmente de frente, con excelente iluminación, sin ningún tipo de accesorio como gafas de sol o gorros y donde se pudiera observar con facilidad el color de ojos, de piel y de cabelo. Se realizaron las pruebas con 15 imágenes por persona, tomando como referencia las etiquetas nombradas en la prueba 2 de reconocimiento facial.

Segunda prueba:

Por otro lado, en la prueba número 2 se realizaron pruebas más reales, en donde se seleccionaron al azar las imágenes de test, tal como se presentó en la prueba 2 de reconocimiento facial se probaron los algoritmos con 50 imágenes diferentes por persona.

6.3.1. Reconocimiento por color de piel

Se utilizó el algoritmo de Segmentación de color por histograma y agrupamiento k-medias, presentado en el marco teórico, el cual tiene

como fin realizar un agrupamiento de pixeles para diferenciar los sectores de la imagen en tres grupos: píxeles de fondo, primer plano o persona y piel.

De igual manera el primero paso es convertir la imagen a escala de grises, luego se aplica la función segment_otsu que permite obtener un umbral con el fin de diferenciar la variación entre clases de la imagen, en este caso lo que es piel del resto de la imagen, asimismo con ayuda de los formatos de color HSV y YCrCb y la formula 5.5, se estima el color de piel de la persona. En la figura 6.9 se muestra un ejemplo del procedimiento:

Figura 6.9. Segmentación de color de piel.

También es importante mencionar que para efectos de simplificación se utilizaron las siguientes categorias de color de piel: clara, morena y oscura. Los resultados se pueden observar en la sección 7.3.1. De igual manera, como se menciona anteriormente se realizaron dos pruebas diferentes, sin embargo el procedimiento es el mismo en ambos casos, la única diferencia son las imágenes de prueba.

6.3.2. Reconocimiento por color de cabello

Para el reconocimiento de color de cabello, se utilizó un algoritmo que utiliza una Red neuronal de convolución basada en regiones [40] preentrenda, el conjunto de datos de entrenamiento cuenta con 1050

imagenes subdividadas en siete tipos diferentes de peinados (lacio, ondulado, rizado, ensortijado, trenzas, rastas, corto) con un total de 160 mil epocas. El procedimiento del algoritmo básicamente es el siguiente: toma la imagen de prueba, la pasa por la red (procedimiento explicado en el marco teórico), y se aplica una máscara donde se separa el cabello del resto de la imagen, luego se promedia el RGB de los pixeles y se obtiene el resultado. En la figura 6.10 una imagen de ejemplo:

Figura 6.10. Segmentación de color de cabello.

Se tuvieron en cuenta las siguientes categorías de color de cabello, dependiendo del RGB arrojado: negro, rubio y rojo.

6.3.3. Reconocimiento por color de ojos

El algoritmo utilizado es una Red convolucional en cascada multitarea MTCNN [41] previamente entrenada (explicación en marco teórico), el conjunto de datos cuenta con 35048 imágenes, El modelo está adaptado

de la implementación MTCNN de Facenet. El detector devuelve una lista de objetos, que incluye las zonas principales del rostro, como la nariz, los ojos y la boca, como el área de interés es los ojos, se descartan las otras características, finalmente con ayuda del formato HSV se encuentra el color correspondiente.

Figura 6.11. Reconocimiento de ojos.

Se tuvieron en cuenta las siguientes categorías de color de ojos, dependiendo del HSV arrojado: cafe, verde y azul.

6.3.4. Teorema de Bayes para características soft-biométricas

En la figura 6.12 se presenta un pequeño resumen del procedimiento para integrar todos los algoritmos por caracteristicas soft-biométricas en uno solo, asimismo más adelante se explica a detalle cada una de las fases.

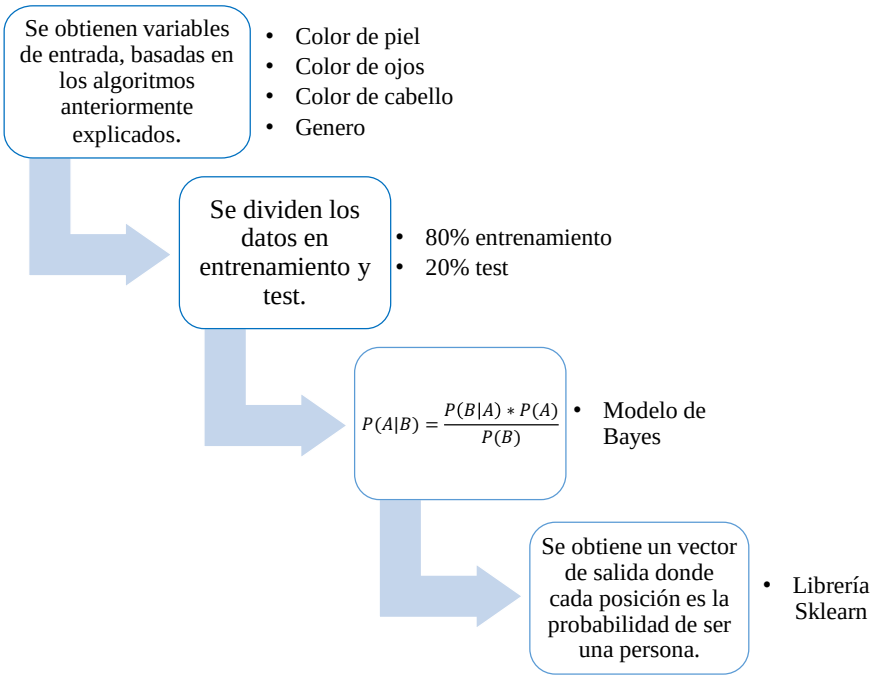

Figura 6.12. Procedimiento reconocimiento por características soft-biométricas.

Una vez obtenidos los resultados de las pruebas de color de piel, ojos, cabello y genero, se procedió a implementar el algoritmo clasificador llamado Gaussian Naive Bayes, que se basa en la ecuación 6 del teorema de Bayes. El procedimiento utilizado fue el siguiente:

Para iniciar, los datos de entrada se tomaron de las imágenes utilizadas para entrenamiento en la sección de reconocimiento de rostro, cabe mencionar que se contaban con 250 imágenes por persona, todas las imágenes fueron procesadas por cada uno de los algoritmos de reconocimiento soft biométrico y de estos se obtuvieron las variables que se presentan en la tabla 6.6:

Tipo	**Variable**
float	Color_Piel_R
float	Color_Piel_G
float	Color_Piel_B
float	Color_Ojos_H
float	Color_Ojos_S
float	Color_Ojos_V
float	Color_Cabello_R
float	Color_Cabello_G
float	Color_Cabello_B
int	Género
int	Persona

Tabla 6.6. Variables teorema de Bayes

Del algoritmo de reconocimiento de color de piel se obtienen las variables de tipo flotante: 'Color_Piel_R', 'Color_Piel_G', 'Color_Piel_B' , que son las componentes RGB del color de piel de la persona detectado en la fotografía.

Del algoritmo de reconocimiento de color de ojos se obtienen las variables de tipo flotante: 'Color_Ojos_H', 'Color_Ojos_S', 'Color_Ojos_V', que son las componentes HSV del color de ojos de la persona detectado en la fotografía.

Del algoritmo de reconocimiento de color de cabello se obtienen las variables de tipo flotante: 'Color_Cabello_R', 'Color_Cabello_G', 'Color_Cabello_B', que son las componentes RGB del color de cabello de la persona detectado en la fotografía.

Del algoritmo de reconocimiento de género se obtiene la variable de tipo entero 'Genero', donde 0 indica masculino y 1 femenino.

Por último se encuentra la variable de tipo entero 'Persona', que indica a qué clase pertenecen los datos, y cuya relación se puede ver a continuación:

- 0 = Abigail_Breslin
- 1 = Andy_Samberg
- 2 = Bryan_Betancur
- 3 = Lina_Plazas
- 4 = Oprah_Winfrey

Posteriormente, con ayuda de la librería Sklearn, se creó el modelo de Bayes, el cual arroja como resultado un vector de probabilidades con tamaño igual a la cantidad de clases o en este caso de personas a identificar. En la tabla 6.7 se muestra un ejemplo:

Personas				
0	1	2	3	4
0,9741	0,0059	0,01038	0,0077	0,0018

Tabla 6.7. Ejemplo de vector de probabilidades

Suponiendo que el vector de probabilidades sea el presentado en la tabla 6.7, quiere decir que según las variables entregadas existe un 97,4 % de probabilidad de que la persona a identificar sea Abigail Breslin, 0,5 % Andy Samberg 0,1 % Bryan Betancur y asi sucesivamente. Se realizaron pruebas con 50 imágenes por persona y se pueden encontrar los resultados

en la siguiente sección.

6.4. Reconocimiento algoritmo multimodal

Luego de los resultados obtenidos en cada uno de los métodos de reconocimiento individuales, se continuó con el tercer objetivo específico, el cual consistía en integrar los algoritmos mencionados anteriormente en un solo desarrollo. Para esto se realizaron tres pruebas que se describen a continuación:

Prueba 1

A partir de los vectores de probabilidades obtenidos en los tres métodos de reconocimiento(facial, voz, soft biométrico), se realizó la prueba más sencilla de todas que fue encontrar el promedio de probabilidades para cada una de las clases.

Prueba 2

Para la segunda prueba se utilizó la función softmax con el fin de que la suma de las probabilidades de los vectores resultantes por lo métodos individuales fuera igual a 1 y de esta manera representar la distribución categórica.

Prueba 3

Para la última prueba se realizó una regresión multivariable como se muestra en la ecuación 6.5:

$$f(x, y, z) = w_1 x + w_2 y + w_3 z \tag{6.5}$$

donde:

x: es la probabilidad de pertenecer a cierta clase por reconocimiento facial.
y: es la probabilidad de pertenecer a cierta clase por reconocimiento de voz.
z: es la probabilidad de pertenecer a cierta clase por reconocimiento de características soft biométricas.

De esta manera se buscaron los pesos que garantizarán un mejor desempeño del algoritmo multimodal, para esto se realizaron pruebas con todas las posibilidades combinaciones del siguiente vector:

$$w = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$$

Las pruebas se volvieron a realizar con el mismo dataset de prueba utilizado en los algoritmos individuales que contaban con 50 imágenes por persona.

6.5. Implementación ROS

La implementación en ROS se hizo por medio de nodos y servicios, el diagrama general del sistema se puede observar en la figura 6.13, además de su correspondiente explicación:

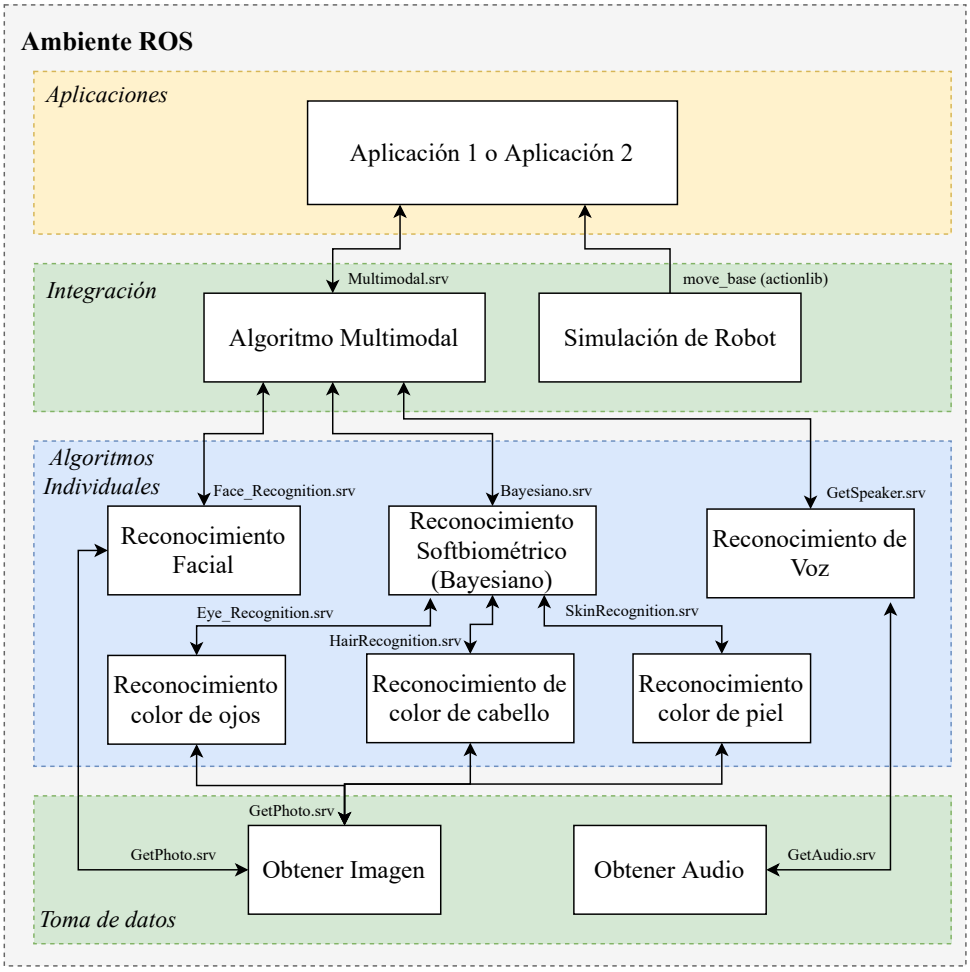

Figura 6.13. Implementación en ROS.

Como se puede observar en la figura 6.13 la implementación en ROS está dividida principalmente en cuatro, a continuación se explica qué contiene cada una de estas capas:

- **Capa de toma de datos**: Se encuentra en la base de la pirámide, ya que dentro de esta capa están los algoritmos que tienen comunicación directa con los dispositivos del computador (cámara y micrófono).
  - Obtener Imagen: Aquí se encuentra el nodo get_photo_node el

cual se encarga de activar la cámara y tomar una foto cuando sea solicitada por medio del servicio GetPhoto.srv.

- Obtener Audio: Aquí se encuentra el nodo get_audio_node el cual se encarga de activar el micrófono y tomar una muestra de audio cuando sea solicitado por medio del servicio GetAudio.srv.

- **Capa de algoritmos individuales**: Esta capa se encarga de ejecutar todos los algoritmos individuales de reconocimiento, así como el algoritmo de simulación del robot.

  - Reconocimiento Facial: Conformado por el nodo face_recognition_node que se encarga de realizar el reconocimiento facial de la fotografía obtenida en get_photo_node y se solicita por medio del servicio FaceRecognition.srv.
  - Reconocimiento de Voz: Conformado por el nodo voice_recognition_node que se encarga de realizar el reconocimiento de voz del audio obtenido en get_audio_node y se solicita por medio del servicio GetSpeaker.srv.
  - Reconocimiento color de piel: Conformado por el nodo skin_recognition_node que se encarga de realizar el reconocimiento de color de piel de la fotografía obtenida en get_photo_node y se solicita por medio del sevicio SkinRecognition.srv.
  - Reconocimiento color de cabello: Conformado por el nodo hair_recognition_node que se encarga de realizar el reconocimiento de color de cabello de la fotografía obtenida en get_photo_node y se solicita por medio del sevicio HairRecognition.srv.
  - Reconocimiento color de ojos: Conformado por el nodo eyes_recognition_node que se encarga de realizar el reconocimiento de color de ojos de la fotografía obtenida en

get_photo_node y se solicita por medio del sevicio Eye_Recognition.srv.

- Reconocimiento Softbiométrico(Bayesiano): Conformado por el nodo bayes_node que se encarga de realizar el clasificador de Naive Bayes y así obtener un solo vector de probabilidades soft biométricas y se solicita por medio del servicio Bayesiano.srv.

▪ **Capa de integración**:

- Algoritmo Multimodal: Está compuesto por el nodo multimodal_node que se encarga de integrar los tres métodos de reconocimiento(facial, voz y características soft-biométricas) en un solo desarrollo, se activa por medio del servicio. Multimodal.srv
- Simulación del robot: La simulación del robot esta conformada por tres archivos tipo Launch:
  - husky_playpen: Se encarga de la ejecución de los nodos necesarios para la carga del mapa y el robot Husky[57] en el entorno de Gazebo.

Figura 6.14. Mapa y Robot en Gazebo

  - view_robot: Con este archivo ìaunch se ejecutan los nodos necesarios para la carga y visualización en Rviz[58] del

robot, con esta herramienta podemos capturar y visualizar los datos obtenidos por los sensores del Husky en el entorno cargado previamente en el anterior launcher.

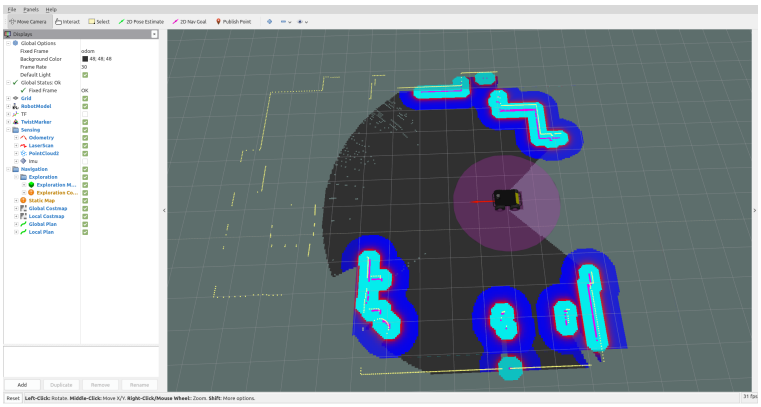

Figura 6.15. Visualización en Rviz

- move_base_mapless_demo: Con la ejecución de este archivo se lanzan los nodos necesarios para la recepción de un punto en el mapa al cual se quiera enviar el robot. Estos nodos realizan la verificacón del punto recibido al igual que la ruta ideal para llegar a este, todo esto con apoyo de los sensores del robot simulados en Rviz.

- **Capa de aplicaciones**: Es donde se encuentran desarrolladas las pruebas 1 y 2, en estos algoritmos se encuentran las estructuras generales de las pruebas, es decir los diálogos entre robot y humano, asimismo se encargan de comunicarse con el algoritmo multimodal y la simulación de los movimientos del robot por medio de servicios.

Capítulo 7

Resultados

A continuación se presentan los resultados de las pruebas mencionadas en la sección 6 de metología.

7.1. Reconocimiento facial

7.1.1. Detección facial

Para la elección del algoritmo de detección facial se utilizaron matrices de confusión (tabla 6.5) obteniendo las métricas *precisión*, *recall*,*acuraccy* (ecuaciones 6.2, 6.3 y 6.4) y tiempo promedio de ejecución para evaluar el desempeño de cada algoritmo.

Para esta prueba se utilizaron 520 fotos, en las cuales aparecian 922 rostros en total, en la sección 6.1.1 de metodología en detección facial se pueden encontrar más detalles de la prueba realizada. De igual manera, En la tabla 7.1 y 7.2 se muestran los resultados obtenidos para los distintos algoritmos de detección facial:

	VP	FN	FP	VN
HOG	704	217	1	14
Viola Jones	691	228	3	14
OpenFace (Hog y Svm)	839	77	6	14

Tabla 7.1. Resultados generales de detección facial

	Precision	Recall	Accuracy	TPD
HOG	99.8581 %	76.43865 %	76.7094 %	0,03S
Viola Jones	99.5677 %	75.1904 %	75.32051 %	0,0318S
OpenFace (Hog y Svm)	99.2899 %	91.5938 %	91.13247 %	0,109S

Tabla 7.2. Métricas de evaluación generales para detección facial

Siendo TPD el tiempo promedio de detección por foto

En los resultados mostrados en la tabla 7.2 se incluía tanto personas adultas como niños y bebés, sin embargo en las tablas 7.3, 7.4, 7.5 y 7.6 se presentan los resultados por separado:

Adultos			
	VP	FN	FP
HOG	597	143	1
Viola Jones	560	178	3
OpenFace (Hog y Svm)	678	59	5

Tabla 7.3. Resultados adultos detección facial

Adultos			
	Precision	**Recall**	**Accuracy**
HOG	99.8327 %	80.6756 %	80.9271 %
Viola Jones	99.4671 %	75.8807 %	76.0264 %
OpenFace (Hog y Svm)	99.4134 %	91.9945 %	91.6556 %

Tabla 7.4. Métricas de evaluación adultos detección facial

Niños			
	VP	**FN**	**FP**
HOG	106	74	0
Viola Jones	131	50	0
OpenFace (Hog y Svm)	161	18	2

Tabla 7.5. Resultados niños detección facial

Niños			
	Precision	**Recall**	**Accuracy**
HOG	100 %	59.1160 %	62.051 %
Viola Jones	100 %	72.3756 %	74.3589 %
OpenFace (Hog y Svm)	98.7730 %	89.9441 %	89.7435 %

Tabla 7.6. Métricas de evaluación niños detección facial

Cabe aclarar que en los resultados por categoría, no se muestran los Verdaderos Negativos (VN) puesto que para las pruebas de dicha métrica, se testea con imágenes sin ningún rostro, por lo cual el resultado no va a depender de la categoría adulto o niño. De igual manera en la figura 7.1 se muestra un ejemplo para complementar la información proporcionada en la tabla 7.6, respecto a la detección en niños:

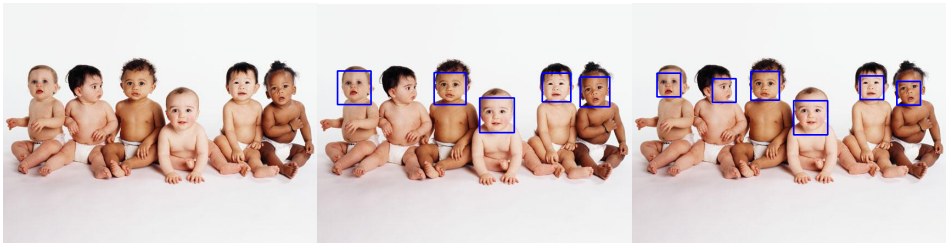

Figura 7.1. Detección facial Viola Jones, Hog, Hog y SVM.

Finalmente el algoritmo que se eligió fue el de openface(Hog y SVM), en la sección 8 de ánalisis de resultados se explica porque se tomó esta decisión.

7.1.2. Reconocimiento facial

Una vez elegido el algoritmo de detección facial (Svm y Hog) con base en los resultados presentados en el inciso anterior, se procedió a realizar las pruebas de los algoritmos de re-identificación o reconocimiento facial, se probaron en total cuatro: Svm, Radial Svm, Grid search Svm y Knn utilizando matrices de confusión (tabla 6.5) obteniendo las métricas precisión, recall, acuraccy (ecuaciones 6.2, 6.3 y 6.4) y tiempo promedio de ejecución para evaluar el desempeño de cada algoritmo.

Para esta prueba se utilizaron 100 imagenes de prueba con 244 rostros incluyendo personas desconocidas, en la sección 6.1.2 de metodologia en reconocimiento facial, se puede encontrar más detalle de la prueba realizada. De igual manera, se utilizaron las ecuaciones 6.2,6.3 y 6.4 para evaluar los algoritmos, obteniendo los siguientes resultados:

Primera prueba:

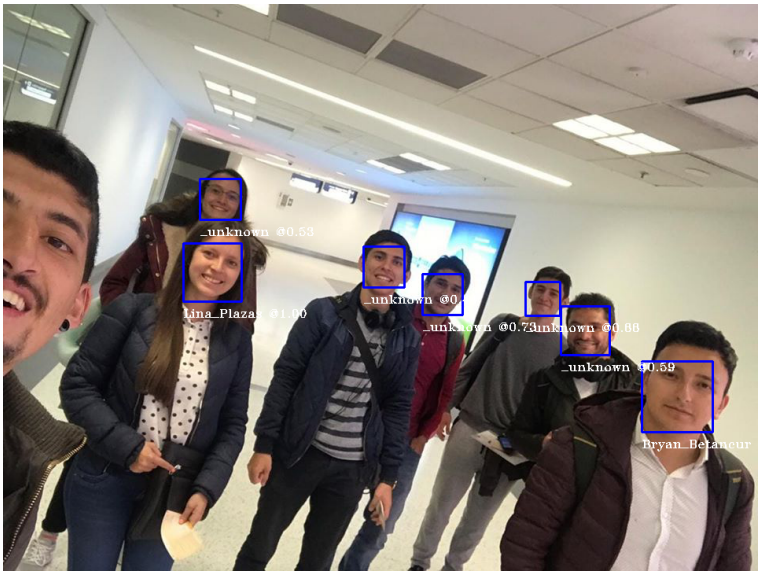

Figura 7.2. Ejemplo reconocimiento facial.

	VP	FN	FP	VN
RadialSvm	103	16	8	117
LinearSvm	110	9	35	90
GridSearchSvm	107	12	38	87
KNN	116	0	64	64

Tabla 7.7. Resultados prueba 1 de reconocimiento facial

	Precision	Recall	Accuracy	TPD
RadialSvm	92.7927 %	86.5546 %	90.1639 %	0,489S
LinearSvm	75.8620 %	92.4369 %	81.9672 %	0.4890S
GridSearchSvm	73.7931 %	89.9159 %	79.5081 %	0,484S
KNN	64.4444 %	100 %	73.7704 %	0,488S

Tabla 7.8. Resultados métricas evaluativas para la prueba 1 de reconocimiento facial.

Siendo TPD el tiempo promedio de reconocimiento por foto
Finalmente se eligio el algoritmo Radial SVM, en la sección 8 de análisis de resultados, se explica porque se tomo está decisión.

Segunda prueba:

Como se menciona en la metodología, para esta prueba únicamente se tuvo en cuenta el algoritmo con mejor desempeño en la prueba 1, es decir el Radial Svm, de igual manera solo se tomaron imágenes de prueba en las que aparecía una sola persona, en las tablas 7.9 y 7.10 se muestran los resultados:

	VP	**FN**	**FP**	**VN**
RadialSvm	237	12	6	45

Tabla 7.9. Resultados métricas evaluativas para la prueba 1 de reconocimiento facial

	Precision	**Recall**	**Accuracy**	**TPF**
RadialSvm	97 %	95.1 %	94.3 %	0.475

Tabla 7.10. Resultados métricas evaluativas prueba 2 reconocimiento facial

Adicionalmente se presenta en la tabla 7.6 la cantidad de aciertos y errores por clase:

	Acertado	No acertado
Abigail Breslin	47	3
Andy Samberg	47	3
Bryan Betancur	44	6
Lina Plazas	50	0
Oprah Winfrey	46	4
Total	234	16
Porcentaje Acierto	**93.6 %**	

Tabla 7.11. Resultados por clases reconocimiento facial

7.2. Reconocimiento de voz

7.2.1. Entrenamiento e identificaci´on del hablante con MFCC-GMM

Para el entrenamiento del algoritmo MFCC-GMM se escogieron 60 hablantes divididos en partes iguales dependiendo si era niño, hombre o mujer. El resultados se muestra en la tabla 7.12:

	Acertado	No acertado
Hombres	609	1
Mujeres	462	0
Niños	30	36
Total	1001	37

Tabla 7.12. Resultados por clases reconocimiento facial

Se realizaron pruebas de este algoritmo entrenando 5 personas y probando con 50 audios de cada persona y a su vez variando la cantidad de audios de entrenamiento. Estos entrenamientos son resumidos en la tabla 7.13 en la que se evidencio que el algortmo MFCC-GMM implementado obtiene un porcentaje de acierto mayor cuando el entrenamiento se hace con aproximadamente 20 audios de alrededor de 12 segundos por cada

hablante, de igual manera se puede notar que se obtuvo un porcentaje de acierto de 95.4 %, resultado acorde a lo indagado en la bibliografia resumida en la tabla 6.3. Tambien se realizarón mediciones del tiempo que se demora el algoritmo realizando el reconcimiento en audios de 4-5 segundos, teniendo como resultado tiempos de 100-200 milisegundos.

MFCC-GMM			
Audios de entrenamiento por persona	**Audios acertados**	**Audios no acertados**	**Porcentaje de acierto**
1	173	77	69.2 %
5	198	52	79.2 %
10	228	22	91.2 %
20	236	14	95.4 %
30	236	14	95.4 %
40	236	14	95.4 %

Tabla 7.13. Resultados MFCC-GMM variando el número de audios de entrenamiento por persona

7.2.2. Entrenamiento e identificación del hablante con MFCC-ResNet

Se realizó el entrenamiento para el algoritmo MFCC-ResNet modificando el optimizador y el porcentaje de validación para los datos de entrenamiento como se puede observar en la tablas 7.13, 7.14 y 7.15 se obtuvo mejores resultados con un porcentaje de validación del 30 % en todos los casos. El mejor resultado obtenido se dió con el optimizador ADAGRAD quien tuvo un porcentaje de acierto del 94 % seguido por el optimizador SGD con 92.4 % y el optimizador ADAM con el 91.6 %. Resultados coherentes a los investigados en el articulo "Performance Comparison of Phoneme Modeling using MFCC and IHCC Features with various Optimization Algorithms for Neural Network Architectures"[6] en el que se concluye que para la representación del habla por medio de la MFCC el mejor optimizador es el

ADAGRAD.
Tal como en el algoritmo anterior se realizaron mediciones del tiempo que se demora el algoritmo realizando el reconocimiento en audios de 4-5 segundos, teniendo como resultado tiempos de 200-300 milisegundos.

MFCC-ResNet ADAGRAD			
Porcentaje de validación	**Audios acertados**	**Audios no acertados**	**Porcentaje de acierto**
10	134	166	53.60 %
20	206	44	82.39 %
30	235	30	94.00 %

Tabla 7.14. Resultados MFCC-ResNet ADAM variando el porcentaje de validación

MFCC-ResNet ADAM			
Porcentaje de validación	**Audios acertados**	**Audios no acertados**	**Porcentaje de acierto**
10	134	166	53.60 %
20	197	53	78.80 %
30	229	21	91.60 %

Tabla 7.15. Resultados MFCC-ResNet SGD variando el porcentaje de validación

MFCC-ResNet SGD			
Porcentaje de validación	**Audios acertados**	**Audios no acertados**	**Porcentaje de acierto**
10	116	34	46.40 %
20	218	32	87.20 %
30	231	19	92.40 %

Tabla 7.16. Resultados MFCC-ResNet ADAGRAD variando el porcentaje de validación

7.2.3. Tiempos de entrenamiento

Para entrenar los algoritmos MFCC-GMM y MFCC-ResNet se utilizó una máquina con procesador Ryzen 9 3900x con una GPU Nvidia Rtx 2070 Super, y se tomó el tiempo que gastó entrenando 80 audios cada uno de los algoritmos, en la tabla 7.17 se pueden observar los resultados que muestra el tiempo que tardó cada algoritmo en crear el modelo o la clase para su posterior reconocimiento. Se puede evidenciar un tiempo de menos de la mitad para realizar el entrenamiento con el algoritmo MFCC-GMM.

	Número de audios	**Tiempo de entrenamiento**
MFCC-GMM	80	7,15 seg
MFCC-ResNet	80	20,4 seg

Tabla 7.17. Comparación tiempo de entrenamiento

7.2.4. Resumen algoritmos de reconocimiento por voz

En tabla 7.18 se resumen los algoritmos implementados que mejores resultados obtuvieron para el reconocimiento de voz.

Resumen algoritmos		
Algoritmo	**Velocidad de reconocimiento**	**Porcentaje de acierto**
MFCC-GMM	100-200 ms	95.4 %
MFCC-ResNet SGD	200-300 ms	87.20 %
MFCC-ResNet ADAM	200-300 ms	92.40 %
MFCC-ResNet ADAGRAD	200-300 ms	94 %

Tabla 7.18. Resultados MFCC-ResNet ADAGRAD variando el porcentaje de validación

7.3. Reconocimiento por características soft biométricas

En esta sección se presentan los resultados de las pruebas explicadas en el capítulo de desarrollo metodológico, las características soft-biométricas que se tuvieron en cuenta en el presente proyecto fueron color de piel, color de ojos y color de cabello. Asimismo, posteriormente se utiliza un clasificador Bayesiano para lograr integrar los tres algoritmos mencionados anteriormente en uno solo.

7.3.1. Reconocimiento por color de piel

Prueba 1

Como se mencionó en la sección de metodología, se utilizó el algoritmo de Segmentación de color por histograma y agrupamiento k-medias y se tuvo en cuenta tres clases para agrupar el color de piel: clara, morena y oscura, en la tabla 7.19 se pueden ver los resultados y en los anexos algunos ejemplos de la implementación:

	Acertado	**No acertado**
Piel clara	40	5
Piel morena	9	6
Piel oscura	9	6
Total	58	17
Porcentaje Acierto	**77.3 %**	

Tabla 7.19. Resultados prueba 1 reconocimiento color de piel

7.3.2. Prueba 2

Bajo el mismo procedimiento anterior, con la única diferencia que se utilizaron 50 fotos por persona, eligiendo las imágenes al azar, se encontraron los resultados presentados en la tabla 7.20:

	Acertado	No acertado
Piel clara	97	53
Piel morena	25	25
Piel oscura	23	27
Total	145	105
Porcentaje Acierto	**58 %**	

Tabla 7.20. Resultados prueba 2 reconocimiento color de piel

7.3.3. Reconocimiento por color de cabello

Prueba 1

Como se mencionó en la sección de metodología,se implentó un algoritmo que utiliza una Red neuronal de convolución basada en regiones y se tuvo en cuenta tres clases para agrupar el color de cabello: rubio, negro y rojo, en la tabla 7.21 se pueden ver los resultados y en los anexos algunos ejemplos de la implementación:

	Acertado	No acertado
Cabello rubio	15	0
Cabello negro	54	6
Total	69	6
Porcentaje Acierto	**92 %**	

Tabla 7.21. Resultados prueba 1 reconocimiento color de cabello

Prueba 2

Bajo el mismo procedimiento anterior, con la única diferencia que se utilizaron 50 fotos por persona, eligiendo las imágenes al azar, se encontraron los resultados presentados en la tabla 7.22:

	Acertado	**No acertado**
Cabello rubio	44	6
Cabello negro	151	49
Total	195	55
Porcentaje Acierto	**78 %**	

Tabla 7.22. Resultados prueba 2 reconocimiento color de cabello

7.3.4. Reconocimiento por color de ojos

Prueba 1

Como se mencionó en la sección de metodología, se utilizo un algoritmo que utiliza una Red convolucional en cascada multitarea MTCNN y se tuvo en cuenta tres clases para agrupar el color de ojos: azul, café y verde, en la tabla 7.23 se pueden ver los resultados y en los anexos algunos ejemplos de la implementación:

	Acertado	**No acertado**
Ojos café	15	0
Ojos azul	41	19
Total	56	19
Porcentaje Acierto	**74.6 %**	

Tabla 7.23. Resultados prueba 2 reconocimiento color de ojos

Prueba 2

Bajo el mismo procedimiento anterior, con la única diferencia que se utilizaron 50 fotos por persona, eligiendo las imágenes al azar, se encontraron los resultados presentados en la tabla 7.24:

	Acertado	**No acertado**
Ojos cafe	139	61
Ojos azul	31	19
Total	170	80
Porcentaje Acierto	**68 %**	

Tabla 7.24. Resultados prueba 2 reconocimiento color de ojos

Teorema de Bayes características softbiométricas

En la tabla 7.25 se presentan los resultados obtenidos luego de implementar el algoritmo de clasificación Gaussian Naive Bayes:

	Acertado	**No acertado**
Abigail Breslin	38	12
Andy Samberg	32	18
Bryan Betancur	33	17
Lina Plazas	27	23
Oprah Winfrey	35	15
Total	165	85
Porcentaje Acierto	**66 %**	

Tabla 7.25. Resultados características soft-biométricas

7.3.5. Reconocimiento por algoritmo multimodal

Para la integración de los algoritmos anteriormente presentados, se realizaron tres pruebas, a continuación se presentan los resultados:

Prueba 1

Obtenidos los resultados provenientes del reconocimiento facial, reconocimiento por voz y reconocimiento por características soft-biométricos donde se tienen tres vectores de cinco posiciones con la probabilidad de ser una persona basado en imágenes y audios, la primera prueba que se realizó fue encontrar el promedio de cada una de las clases, los resultados se presentan en la tabla 7.26:

	Acertado	**No acertado**
Abigail Breslin	48	2
Andy Samberg	46	4
Bryan Betancur	44	6
Lina Plazas	50	0
Oprah Winfrey	47	3
Total	235	15
Porcentaje Acierto	**94 %**	

Tabla 7.26. Resultados prueba 1 algoritmo multimodal

Prueba 2

Luego de haber realizado la prueba con el promedio de los datos, se probó utilizando la función softmax en los vectores de resultados individuales, de esta manera la suma de las posiciones se normalizaba a 1. En la tabla 7.27 se muestran los resultados:

	Acertado	No acertado
Abigail Breslin	48	2
Andy Samberg	46	4
Bryan Betancur	44	6
Lina Plazas	50	0
Oprah Winfrey	47	3
Total	235	15
Porcentaje Acierto	**94 %**	

Tabla 7.27. Resultados prueba 2 algoritmo multimodal

Como se puede observar los resultados son los mismos tanto con el promedio como utilizando softmax.

Prueba 3

Como se explicó en la metodología, se realizó un barrido con las posibles combinaciones de pesos entre 0,1 hasta 1, con el fin de elegir la que presentará más número de aciertos.En la tabla 7.28 se presentan las mejores 10 combinaciones:

peso 1	peso 2	peso 3	Cantidad de aciertos
0.5	0.8	0.3	244
0.3	0.7	0.2	243
0.1	0.3	0.1	242
0.1	0.7	0.1	241
0.5	0.9	0.3	241
0.1	0.6	0.1	240
0.1	1	0.1	238
0.1	0.8	0.1	238
0.2	0.5	0.1	238

Tabla 7.28. Elección de pesos algoritmo multimodal

Como se puede observar la mejor combinación de pesos es w1=0.5, w2=0.8 y w3=0.3. De esta manera los resultados obtenidos utilizando la ecuación 6.5 se pueden observar en la tabla 7.29:

	Acertado	**No acertado**
Abigail Breslin	48	2
Andy Samberg	48	2
Bryan Betancur	50	0
Lina Plazas	50	0
Oprah Winfrey	48	2
Total	244	6
Porcentaje Acierto	**97.6 %**	

Tabla 7.29. Resultados prueba 3 algoritmo multimodal

Finalmente en la tabla 7.30, se presenta una comparación entre los algoritmos implementados individualmente y el algoritmo multimodal:

Algoritmo	**Porcentaje de acierto**
Reconocimiento facial, Radial SVM(tabla 7.6)	93.6 %
Reconocimiento de voz, MFCC-GMM(tabla 7.13)	94.3 %
Reconocimiento color de piel, segmentación por histograma y k-mdias(Tabla 7.20)	58 %
Reconocimiento color de ojos, CNN basada en regiones (Tabla 7.24)	68 %
Reconocimiento color de cabello, MTCNN (Tabla 7.22)	78 %
Reconocimiento algoritmo multimodal (Tabla 7.30)	97.6 %

Tabla 7.30. Métricas de evaluación adultos detección facial

En la tabla 7.30 se puede observar una comparación entre los porcentajes de acierto de los algoritmos implementados individualmente como el multimodal, cabe aclarar que en todas las pruebas se utilizaron las

mismas imágenes, fueron en total 50 por persona.

7.3.6. Implementación en ROS

En la figura 6.13 de la sección de metodología se puede observar todo el sistema implementado en ROS y la correspondiente explicación de los nodos y servicios. De igual manera,para las pruebas en donde se evalúa si el robot cumple ordenes personalizadas en un ambiente familiar, se tomó como base la competencia Robocup@home 2019 [56]. A continuación se presenta la explicación de cada una de las pruebas:

Primera prueba

Esta prueba tiene como objetivo verificar que el robot sea capaz de tomar y entregar un pedido de bebida correctamente reconociendo a la persona que lo solicitó. Cabe aclarar que las personas de la prueba están previamente entrenadas.En la figura 7.3 se puede observar un resumen de la prueba:

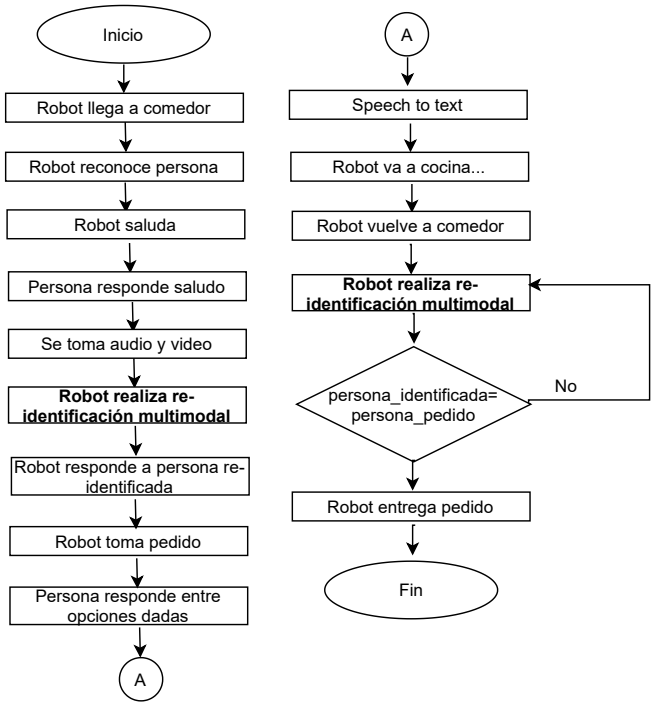

Figura 7.3. Prueba 1 en ambiente familia(fuente autor)

A continuación se explica detalladamente cada uno de los pasos de la prueba:

- La prueba inicia cuando el robot llega al comedor (simulación Gazebo).
- Seguidamente el robot valida que haya una persona frente a él (para este caso se simula con una acción manual, oprimir tecla).
- Luego de validar que haya una persona frente a él, el robot saluda.
- La persona responde el saludo y mientras lo hace se toma muestra de audio y video para su posterior análisis.
- La imagen de la persona se envía a los nodos de reconocimiento facial y reconocimiento de características soft-biométricas, mientras que el audio lo recibe el nodo de reconocimiento de voz.

- Posteriormente, se activa el nodo de re-identificación multimodal y arroja el resultado de la persona identificada, con base en lo que recibe de los diferentes nodos de reconocimiento individual(facial,voz,características soft-biométricas).
- El robot responde haciendo énfasis en el nombre de la persona que reconoció el algoritmo multimodal y le solicita que elija una bebida, con base en tres opciones dadas.
- La persona responde la bebida que desea ordenar.
- Por medio de un algoritmo de habla a texto, se analiza qué bebida eligió la persona.
- Una vez que el robot ha rectificado que entendió bien la bebida, se dirige a la cocina (simulación Gazebo).
- Tiempo después el robot regresa al comedor con la orden solicitada (simulación Gazebo).
- El robot se sitúa frente a la primera persona que encuentra y lo saluda nuevamente.
- La persona responde y mientras lo hace se toma muestra de audio y vídeo para su posterior análisis.
- Se vuelve a realizar re-identificación multimodal.
- Finalmente si la persona identificada, es la misma que solicitó el pedido, el robot le entrega su orden, si no es así, el robot sigue buscando a la persona.

Segunda prueba

Esta prueba tiene como objetivo verificar que el robot sea capaz de cumplir una orden básica, como lo es entregar un objeto a una persona especifica. Cabe aclarar que las personas de la prueba están previamente entrenadas. En la figura 7.4 se puede observar un resumen de la prueba:

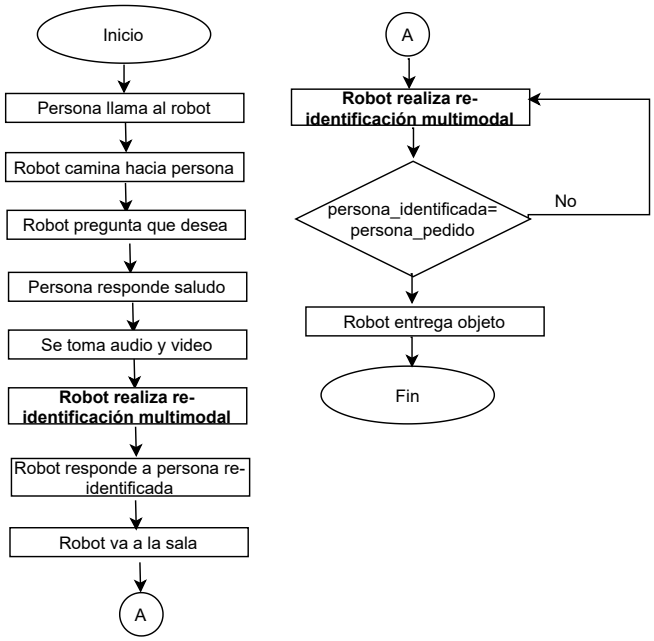

Figura 7.4. Prueba 2 en ambiente familia(fuente autor)

A continuación se explica detalladamente cada uno de los pasos de la prueba:

- La prueba inicia cuando la persona llama al robot.
- El robot se dirige hacia la persona (simulación Gazebo).
- El robot saluda a la persona.
- La persona responde saludo y le solicita un requerimiento a una persona específica, por ejemplo: "Llévale estas llaves a Bryan", mientras tanto se toma muestra de audio y vídeo para su posterior análisis.
- La imagen de la persona se envía a los nodos de reconocimiento facial y reconocimiento de características soft-biométricas, mientras que el audio lo recibe el nodo de reconocimiento de voz.

- Posteriormente, se activa el nodo de re-identifiación multimodal y arroja el resultado de la persona identificada, con base en lo que recibe de los diferentes nodos de reconocimiento individual(facial,voz,características soft biométricas).
- El robot responde haciendo énfasis en el nombre de la persona que reconoció el algoritmo multimodal.
- El robot se dirige a la sala (simulación Gazebo).
- El robot se sitúa frente a la primera persona que encuentra y lo saluda.
- La persona responde y mientras lo hace se toma muestra de audio y vídeo para su posterior análisis
- Se vuelve a realizar re-identificación multimodal.
- Finalmente, si el nombre de la persona identificada corresponde al nombre que le dieron inicialmente, en este ejemplo: Si la persona identificada es igual a Bryan, el robot le entrega su orden, si no es así,el robot sigue buscando a la persona

Resultados Finales

Con base en las pruebas explicadas anteriormente, se realizaron 50 simulaciones en tiempo real de cada una de ellas, en las que se evaluó si el algoritmo realizaba correctamente la re-identificación multimodal. En las figuras 7.5 y 7.6 se puede observar la probabilidad obtenida para cada una de las clases o personas entrenadas con el algoritmo multimodal.

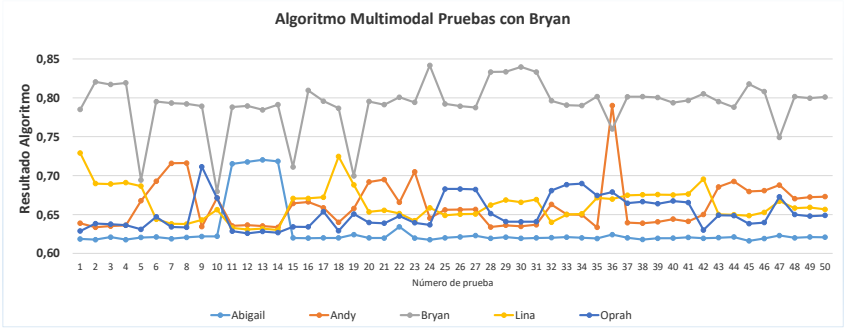

Figura 7.5. Resultados pruebas Bryan Betancur (Fuente Autor)

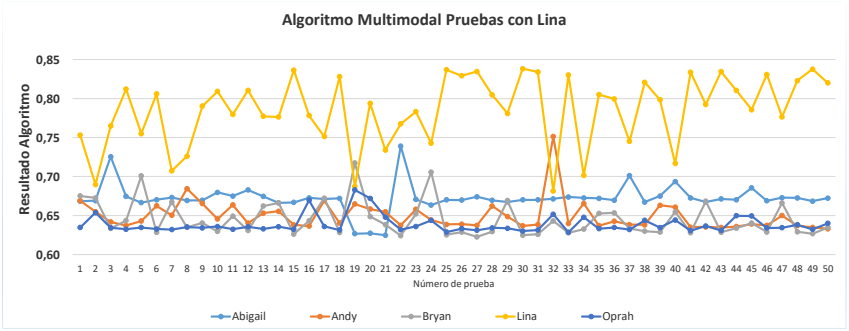

Figura 7.6. Resultados pruebas Lina Plazas (Fuente Autor)

En la figura 7.5 se puede observar que el algoritmo supera en la mayoría de casos la probabilidad de 0.75. De igual manera, en las simulaciones realizadas con la clase Bryan_Betancur se obtuvo un falso reconocimiento. Para las pruebas realizadas con la clase Lina_Plazas (Figura 7.6) se observa que en dos ocasiones el algoritmo reconoce equivocadamente a las

clases Bryan_Betancur y Andy_Samberg. Estos resultados son acordes a los presentados en la Tabla 7.30 en la que se muestra que el porcentaje de acierto para el algoritmo multimodal es de 97.6 %

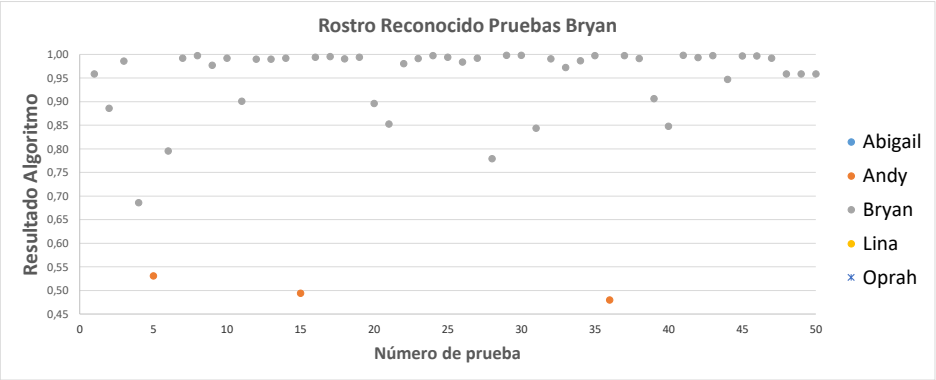

Figura 7.7. Resultados de Rostro Bryan Betancur (fuente autor)

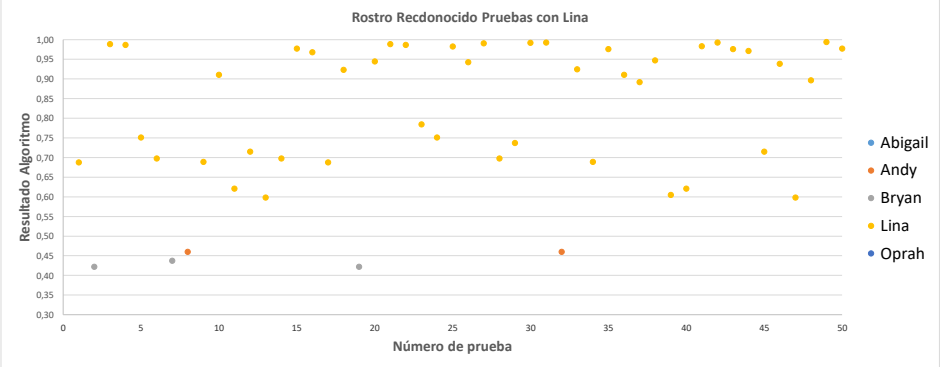

Figura 7.8. Resultados de Rostro Lina Plazas (fuente autor)

En las figuras 7.7 y 7.8 se presentan los resultados obtenidos para el algoritmo de reconocimiento de rostro para las simulaciones mencionadas con anterioridad. En la figura 7.7 muestra que para el caso de la clase Bryan_Betancur el algoritmo tuvo 47 aciertos y para la clase Lina_Plazas el algoritmo reconoció 45 veces correctamente. Esto representa un porcentaje de acierto de 90 % y 94 % respectivamente.

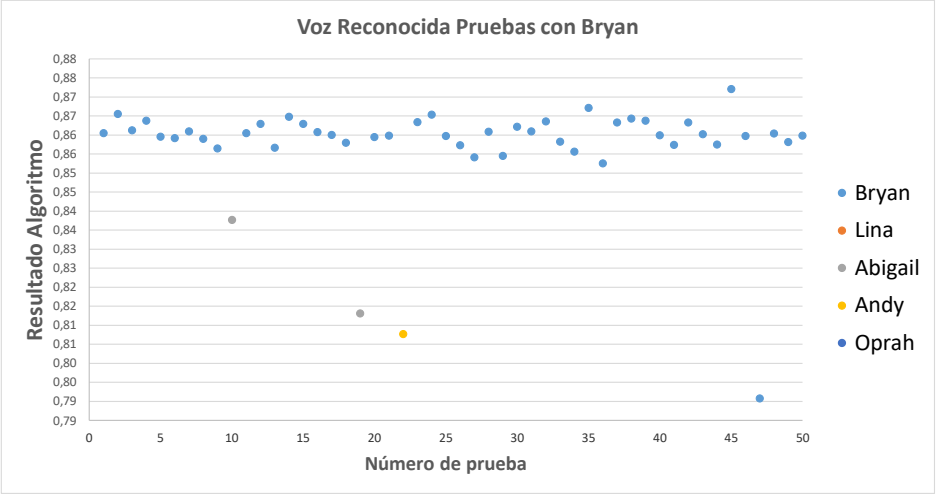

Figura 7.9. Resultados de Voz Bryan Betancur (fuente autor)

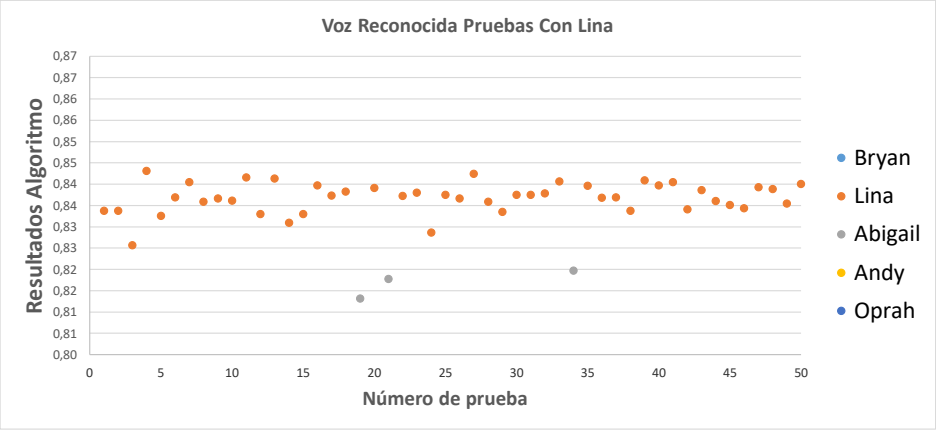

Figura 7.10. Resultados de Voz Lina Plazas (fuente autor)

En las figuras 7.9 y 7.10 se reflejan los resultados para el algoritmo de reconocimiento de voz implementado en las 50 simulaciones explicadas anteriormente. De ellas se puede inferir que para cada una de las clases se obtuvo un porcentaje de acierto del 94 %.

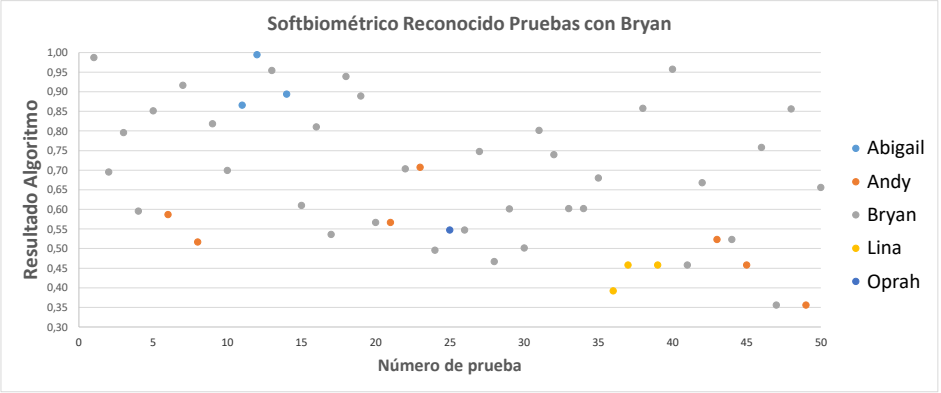

Figura 7.11. Resultados Soft-biométricos Bryan Betancur (fuente autor)

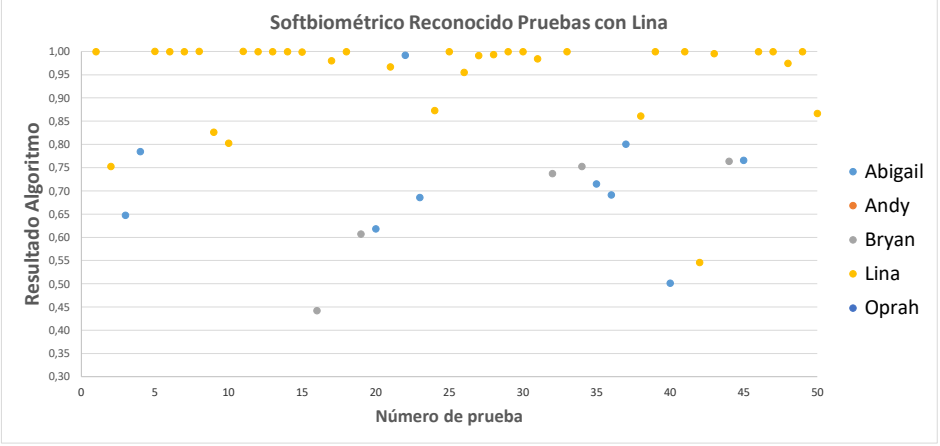

Figura 7.12. Resultados Soft-biométricos Lina Plazas (fuente autor)

En las figuras 7.11 y 7.12 se dan a conocer los resultados para el algoritmo implementado para el reconocimiento de las características soft-biometricas en el que se obtuvo 35 aciertos para cada uno de los casos, siendo este el algoritmo con menor porcentaje de acierto.

Capítulo 8

Análisis de Resultados

Luego de las pruebas realizadas en cuanto a detección facial, las cuales se pueden evidenciar desde la tabla 7.1 hasta la 7.6, se eligió el algoritmo OpenFace (Hog y Svm), ya que como se observa en la tabla 7.4, a pesar de que el algoritmo en cuanto a tiempo de detección es más lento que los demás, tanto la precisión, como el recall y el accuracy es mayor. Asimismo, en la tabla 7.6 se puede observar que las métricas de evaluación del algoritmo OpenFace son mejores para imágenes en las que aparecen niños y bebés en comparación a los algoritmos Hog y Viola Jones.

En cuanto a reconocimiento facial, en la tabla 7.8 se puede observar que todos los algoritmos presentan un tiempo promedio de reconocimiento muy similar, sin embargo en cuanto a precision y accuracy el algoritmo que presenta un mejor desempeño es el Radial SVM por lo cual se decidió elegirlo para la integración multimodal.

Asimismo para la elección del algoritmo que realiza el reconocimiento de hablante se tuvo en cuenta la precisión y la velocidad de reconocimiento, como se puede notar en la tabla 7.17 el algoritmo que mejor precisión y velocidad obtuvo fue el MFCC-GMM que a su vez presenta una facilidad mayor a la hora de realizar el entrenamiento de un nuevo hablante, por

ende fue el algoritmo seleccionado para la implementación final con el algoritmo multimodal.

En la tabla 7.29 se puede observar una comparación entre el algoritmo multimodal y los algoritmos individuales, donde se puede observar que efectivamente al haber una integración de los algoritmos en un único desarrollo, el porcentaje de acierto es mayor.

Conclusiones

En relación a los resultados obtenidos en la sección de detección de rostros, es posible concluir que el algoritmo utilizado en la librería *Openface* presenta con bastante diferencia mejor *accuracy* y *recall* que los algoritmos Hog y Viola Jones, esto se debe a que tiene un menor número de falsos negativos y falsos positivos, lo anterior se puede evidenciar en las tablas 7.1 y 7.2. Asimismo, es posible observar en las pruebas con niños y bebés (tablas 7.5 y 7.6) que a pesar de que se mantiene la afirmación mencionada anteriormente, para los tres algoritmos se presenta una disminución en las métricas de accuracy y recall a comparación de las pruebas realizadas con adultos.

La metodología de entrenamiento por mezcla de modelos Gaussianos es significativamente sencilla y veloz respecto a la empleada con redes neuronales, ya que su diferencia al entrenar el mismo número de audios fue alrededor de 65 % menor en tiempo (tabla 7.18) y no necesita el uso de GPU para aumentar su velocidad, a diferencia del entrenamiento por ResNet, que adicional a esto necesita la implementación de librerías adicionales para un rendimiento óptimo.

A partir de los resultados obtenidos al realizar el entrenamiento se logra evidenciar que el algoritmo MFCC-GMM reduce su precisión al tratar de identificar la voz de un menor de edad en comparación de la voz de un adulto, lo cual se puede evidenciar en la tabla 7.12. Esto se debe a que los audios obtenidos que constituyeron el dataset para el entrenamiento

contenían voces de niños en las que solo se pronunciaban vocales o palabras muy cortas.

Según las tabla 7.13, se puede evidenciar que el algoritmo MFCC-GMM no presenta una mejoría en su porcentaje de acierto a pesar de incrementar el número de audios por persona. Esto se atribuye a un sobre-entrenamiento *(overfit)*[52] por lo tanto, es posible concluir que es necesario encontrar un punto de equilibrio en los datos de entrenamiento, el cual para las pruebas realizadas en el presente proyecto fue de alrededor de 20 audios por persona, llegando a un porcentaje de acierto de 95.3 %.

Ambos algoritmos de identificación del hablante implementados en este trabajo obtuvieron resultados superiores al 94 % de precisión (tabla 7.17), sin embargo para la integración multimodal se eligió el algoritmo que usa los modelos de mezcla Gaussiana, debido a que necesita menos componentes de hardware para funcionar eficazmente.

Al realizar la implementación del clasificador Bayesiano que integraba las variables obtenidas en el reconocimiento por color de cabello, de ojos, de piel y de género, como se puede observar en el cuadro 7.25, no se presenta un porcentaje de acierto muy alto, esto se debe a que efectivamente las características soft-biométricas no tienen la misma capacidad de reconocimiento que por voz y por rostro, sin embargo como se puede evidenciar más adelante en el cuadro 7.56 sirven de apoyo para una identificación más acertada.

Por otro lado, durante la elaboración del proyecto se pudo observar que tanto para el reconocimiento facial como para el reconocimiento por color de ojos, de cabello y de piel, es muy importante la calidad de las imágenes en cuanto a resolución, iluminación y posicionamiento de las personas, tanto en los datos de entrenamiento como en los de prueba, lo cual se evidencio en las pruebas independientes de los algoritmos mencionados y

confirma los problemas mencionados en los artículos [17] y [18] al realizar la re-identificación de personas por un método en particular.

Finalmente, en el tabla 7.29 se puede observar que el algoritmo multimodal alcanza a obtener un porcentaje de acierto del 97.6 % lo cual demuestra que al integrar los algoritmos, efectivamente existe una mejora significativa en el reconocimiento de personas, de igual manera con la implementación en ROS se puede evidenciar una fácil migración del desarrollo a una plataforma robótica.

Capítulo 9

Trabajos Futuros

A continuación, se presentan algunos trabajos futuros que podrían surgir como resultado del presente proyecto:

En primer lugar, se podría realizar la implementación del algoritmo de reconocimiento multimodal en una plataforma robótica como Pepper o Darwin que actualmente se encuentran en el laboratorio de robótica de la Facultad de Ingeniería Electrónica de la Universidad Santo Tomás, para esto se puede tomar como base los nodos implementados en ROS que se presentan en la figura 6.14, lo anterior con el fin de lograr una fácil migración del sistema al robot.

De igual manera en un futuro, se podría plantear el desafío de reconocer varias personas al mismo tiempo, como se puede evidenciar en la sección de resultados, el algoritmo de reconocimiento facial ya cuenta con esta funcionalidad, sin embargo el reto se encuentra en el algoritmo de reconocimiento por voz , ya que se debe implementar una nueva funcionalidad que logre separar múltiples voces en un mismo audio para lograr realizar la validación junto al reconocimiento facial.

Por último con el fin de que el sistema sea más robusto, en un futuro se podría en primer lugar entrenar más personas así como lograr identificar

desconocidos, puede ser cambiando el clasificador bayesiano utilizado para integrar las características soft-biométricas o ampliando el dataset con una nueva categoría que represente a distintas personas que no hacen parte del grupo que se desea reconocer.

Apéndice A

Ejemplos caracteristicas soft-biométricas

A continuación se presentan ejemplos de reconocimiento de color de piel, color de cabello y color de ojos:

Figura A.1. Ejemplo 1 resultados reconocimiento color de piel

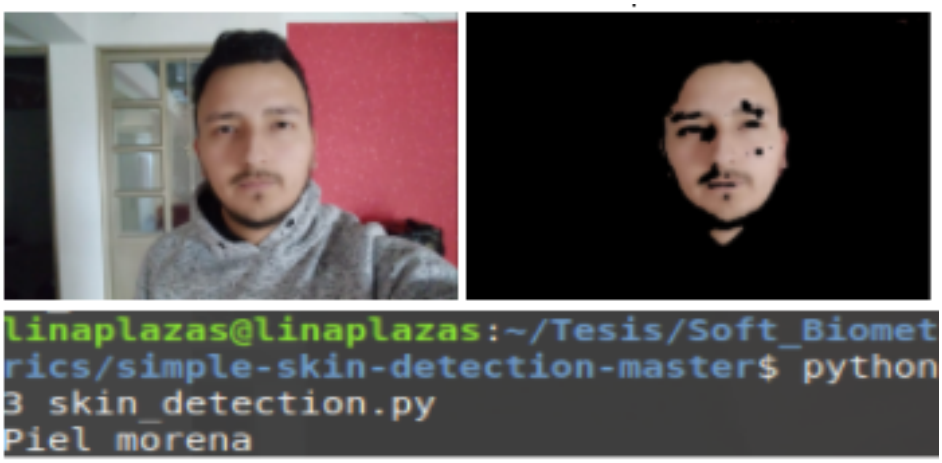

Figura A.2. Ejemplo 2 resultados reconocimiento color de piel

Figura A.3. Ejemplo 2 resultados reconocimiento color de piel

Figura A.4. Ejemplo 1 reconocimiento color de cabello

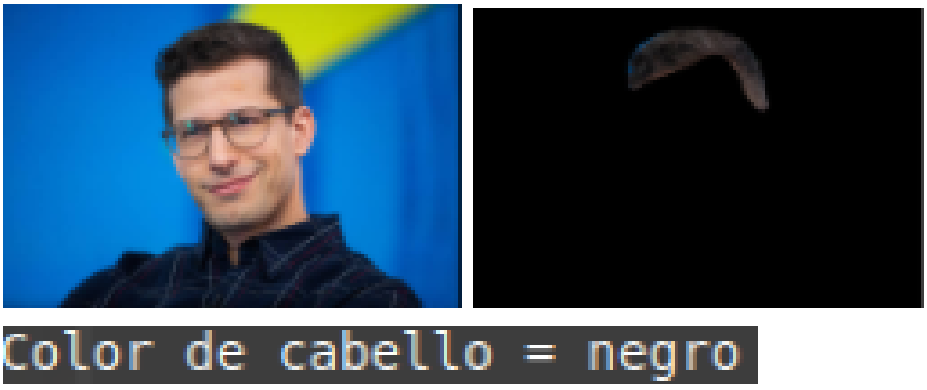

Figura A.5. Ejemplo 2 reconocimiento color de cabello

Figura A.6. Ejemplo 1 reconocimiento color de ojos

Figura A.7. Ejemplo 2 resultado reconocimiento color de ojos

Bibliografía

[1] Alpaydin E. Introduction to machine learning. MIT press; 2020.

[2] Kapusta A, Erickson Z, Clever HM, et al. Personalized collaborative plans for robot-assisted dressing via optimization and simulation. Autonomous Robots. 2019;43(8):2183-2207.

[3] Zhang DD. Automated biometrics: Technologies and systems. Vol 7. Springer Science & Business Media; 2013.

[4] Alejandra MD. Re-identificación de personas a través de sus características soft-biométricas en un entorno multi-cámara de video-vigilancia. Ingeniería, investigación y tecnología. 2016;17(2):257-271.

[5] Norvig P, Russell S. Inteligencia artificial. Editora Campus. 2004;20.

[6] Cong DT, Khoudour L, Achard C, Meurie C, Lezoray O. People re-identification by spectral classification of silhouettes. Signal Process. 2010;90(8):2362-2374.

[7] Salter T, Werry I, Michaud F. Going into the wild in child–robot interaction studies: Issues in social robotic development. Intelligent Service Robotics. 2008;1(2):93-108.

[8] Wechsler H, Phillips JP, Bruce V, Soulie FF, Huang TS. Face recognition: From theory to applications. Vol 163. Springer Science & Business Media; 2012.

[9] Zeng J, Zeng J, Qiu X. Deep learning based forensic face verification in videos. . 2017:77-80.

[10] Rashid RA, Mahalin NH, Sarijari MA, Aziz AAA. Security system using biometric technology: Design and implementation of voice recognition system (VRS). . 2008:898-902.

[11] Weinrich C, Volkhardt M, Gross H. Appearance-based 3d upper-body pose estimation and person re-identification on mobile robots. . 2013:4384-4390.

[12] Barea R, Bergasa LM, López E, Escudero MS, Leon C. Face recognition for social interaction with a personal robotic assistant. . 2005;1:382-385.

[13] Hsu C, Lin C, Su W, Cheung G. Sigan: Siamese generative adversarial network for identity-preserving face hallucination. IEEE Trans Image Process. 2019;28(12):6225-6236.

[14] D'Auria D, Persia F, Bettini F, Helmer S, Siciliano B. Sarri: A smart rapiro robot integrating a framework for automatic high-level surveillance event detection. . 2018:238-241.

[15] Matamoros M, Rascon C, Hart J, et al. RoboCup@Home rules & regulations. .

[16] Kim D, Yoon H, Chi S, Cho Y. Face identification for human robot interaction: Intelligent security system for multi-user working environment on PC. . 2006:617-622.

[17] Akbulut Y, Şengür A, Budak Ü, Ekici S. Deep learning based face liveness detection in videos. . 2017:1-4.

[18] Bae H, Lee H, Lee S. Voice recognition based on adaptive MFCC and deep learning. . 2016:1542-1546.

[19] Chakraborty K, Talele A, Upadhya S. Voice recognition using MFCC algorithm. International Journal of Innovative Research in Advanced Engineering (IJIRAE). 2014;1(10):2349-2163.

[20] Fang X, Gu W, Huang C. A method of skin color identification based on color classification. . 2011;4:2355-2358.

[21] Song A, Hu Q, Ding X, Di X, Song Z. Similar face recognition using the IE-CNN model. IEEE Access. 2020;8:45244-45253.

[22] Sinith MS, Salim A, Sankar KG, Narayanan KS, Soman V. A novel method for text-independent speaker identification using mfcc and gmm. . 2010:292-296.

[23] Połap D, Woźniak M. Voice recognition by neuro-heuristic method. Tsinghua Science and Technology. 2018;24(1):9-17.

[24] Christensen HI, Batzinger T, Bekris K, et al. A roadmap for us robotics: From internet to robotics. Computing Community Consortium. 2009;44.

[25] Siciliano B, Caccavale F, Zwicker E, et al. Euroc-the challenge initiative for european robotics. . 2014:1-7.

[26] Tome P, Fierrez J, Vera-Rodriguez R, Nixon MS. Soft biometrics and their application in person recognition at a distance. IEEE Transactions on information forensics and security. 2014;9(3):464-475.

[27] Shan C, Gong S, McOwan PW. Facial expression recognition based on local binary patterns: A comprehensive study. Image Vision Comput. 2009;27(6):803-816.

[28] Ephraim T, Himmelman T, Siddiqi K. Real-time viola-jones face detection in a web browser. . 2009:321-328.

[29] Suárez EJC. Tutorial sobre máquinas de vectores soporte (sVM). Tutorial sobre Máquinas de Vectores Soporte (SVM). 2014:1-12.

[30] Chandrasekhar AM, Raghuveer K. An effective technique for intrusion detection using neuro -fuzzy and radial svm classifier. In: Computer networks & communications (NetCom). Springer; 2013:499-507.

[31] Olabe XB. Redes neuronales artificiales y sus aplicaciones. Publicaciones de la Escuela de Ingenieros. 1998.

[32] Gazalba I, Reza NGI. Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification. . 2017:294-298.

[33] Muda L, Begam M, Elamvazuthi I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083. 2010.

[34] Eskidere Ö, Gürhanlı A. Voice disorder classification based on multitaper mel frequency cepstral coefficients features. Computational and mathematical methods in medicine. 2015;2015.

[35] Gupta A, Gupta H. Applications of MFCC and vector quantization in speaker recognition. . 2013:170-173.

[36] Huang T, Peng H, Zhang K. Model selection for gaussian mixture models. Statistica Sinica. 2017:147-169.

[37] Buza E, Akagic A, Omanovic S. Skin detection based on image color segmentation with histogram and k-means clustering. . 2017:1181-1186.

[38] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. . 2017:2961-2969.

[39] Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett. 2016;23(10):1499-1503.

[40] Asafu-Adjei JK, Betensky RA. A pairwise naïve bayes approach to bayesian classification. Int J Pat Recognit Artif Intell. 2015;29(07):1550023.

[41] Jang E, Gu S, Poole B. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144. 2016.

[42] Furui S, Kikuchi T, Shinnaka Y, Hori C. Speech-to-text and speech-to-speech summarization of spontaneous speech. IEEE Transactions on Speech and Audio Processing. 2004;12(4):401-408.

[43] ROS.org. https://wiki.ros.org/.

[44] Rahmad C, Asmara RA, Putra D, Dharma I, Darmono H, Muhiqqin I. Comparison of viola-jones haar cascade classifier and histogram of oriented gradients (HOG) for face detection. . 2020;732(1):012038.

[45] Amos B, Ludwiczuk B, Satyanarayanan M. Openface: A general-purpose face recognition library with mobile applications. CMU School of Computer Science. 2016;6(2).

[46] Shinwari AR, Jalali Balooch A, Alariki AA, Abduljalil Abdulhak S. A comparative study of face recognition algorithms under facial expression and illumination. ICACT. Feb 2019:390-394. https://ieeexplore.ieee.org/document/8702002. doi: 10.23919/ICACT.2019.8702002.

[47] Weng Z, Li L, Guo D. Speaker recognition using weighted dynamic MFCC based on GMM. . 2010:285-288.

[48] Shahin I, Nassif AB, Hamsa S. Novel cascaded gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments. Neural Computing and Applications. 2020;32(7):2575-2587.

[49] Kumari TJ, Jayanna HS. Comparison of LPCC and MFCC features and GMM and GMM-UBM modeling for limited data speaker verification. . 2014:1-6.

[50] Thakker M, Vyas S, Ved P, Therese SS. Speaker identification in a multi-speaker environment. In: Information and communication technology for sustainable development. Springer; 2018:239-244.

[51] Amante R, S, Tingzon I. Speaker recognition using mel frequency cepstral coefficients with spectral subband centroids features. .

[52] Reynolds DA, Rose RC. Robust text-independent speaker identification using gaussian mixture speaker models. IEEE transactions on speech and audio processing. 1995;3(1):72-83.

[53] Rozario B, Thomas A, Mathew D. Performance comparison of phoneme modeling using MFCC and IHCC features with various optimization algorithms for neural network architectures. . 2019:240-245.

[54] Kamruzzaman SM, Karim A, Islam M, Haque M. Speaker identification using mfcc-domain support vector machine. arXiv preprint arXiv:1009.4972. 2010.

[55] Mansour A, Lachiri Z. SVM based emotional speaker recognition using MFCC-SDC features. International Journal of Advanced Computer Science and Applications. 2017;8(4):538-544.

[56] Di Nuovo A, Conti D, Trubia G, Buono S, Di Nuovo S. Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability. Robotics. 2018;7(2):25.

[57] Clearpath Robotics. Husky UGV - outdoor field research robot by clearpath. https://clearpathrobotics.com/husky-unmanned ground-vehicle-robot/.

[58] ROS Wiki. rviz. http://wiki.ros.org/rviz.