UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

TESIS DOCTORAL

Estudio de Técnicas de Biometría Blanda en Entornos de Caracterización de Personas. Aplicaciones en Diarización

Programa de Doctorado en Tecnologías de Telecomunicación e Ingeniería Computacional

Pedro Antonio Marín Reyes

Las Palmas de Gran Canaria

2022

D/Dª..... COORDINADOR/A DEL PROGRAMA DE DOCTORADO TECNOLOGÍAS DE TELECOMUNICACIÓN E INGENIERÍA COMPUTACIONAL DE LA UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

INFORMA,

De que la Comisión Académica del Programa de Doctorado, en su sesión de fecha ..... tomó el acuerdo de dar el consentimiento para su tramitación, a la tesis doctoral titulada "Estudio de Técnicas de Biometría Blanda en Entornos de Caracterización de Personas. Aplicaciones en Diarización" presentada por el doctorando D. Pedro Antonio Marín Reyes y dirigida por el Doctor José Javier Lorenzo Navarro.

Y para que así conste, y a efectos de lo previsto en el Artº 11 del Reglamento de Estudios de Doctorado (BOULPGC 04/03/2019) de la Universidad de Las Palmas de Gran Canaria, firmo la presente en Las Palmas de Gran Canaria, a...de.....de dos mil.....

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
ESCUELA DE DOCTORADO

Programa de doctorado Tecnologías de Telecomunicación e Ingeniería Computacional

Título de la Tesis: Estudio de Técnicas de Biometría Blanda en Entornos de Caracterización de Personas. Aplicaciones en Diarización

Tesis Doctoral presentada por D. Pedro Antonio Marín Reyes

Dirigida por el Dr. José Javier Lorenzo Navarro

Codirigida por el Dr. Modesto Fernando Castrillón Santana

Las Palmas de Gran Canaria, a 7 de febrero de 2021

El Director, El Codirector El Doctorando,

(firma) (firma) (firma)

Estudio de Técnicas de Biometría Blanda en Entornos de Caracterización de Personas. Aplicaciones en Diarización

Pedro Antonio Marín Reyes

Programa de Doctorado
Tecnologías de Telecomunicación e Ingeniería Computacional

Directores:
José Javier Lorenzo Navarro
Modesto Fernando Castrillón Santana

LAS PALMAS DE G.C. Febrero 2022

RESUMEN

La biometría blanda representa esas características que permiten identificar de forma aproximada a una persona. Cuantas más características de este tipo se escojan mejor será la caracterización de la persona. Además, si incorporamos técnicas de biometría fuerte, se afianza la estimación resultante del modelado de la persona. Estos factores tienen vital importancia a la hora de identificar/re-identificar a personas. Por un lado, se conoce por identificación cuando el sistema es capaz de determinar quién es la persona dentro de una base de datos de usuarios previamente almacenada. Por otro lado, hablamos de re-identificación cuando el sistema no tiene almacenado la información de los usuarios que se espera identificar dentro de una red de cámaras.

En esta tesis se realiza un análisis sobre diferentes factores que afectan a la caracterización de personas, desde la fase del modelado de la persona hasta la influencia del escenario en estas. En la actualidad existen múltiples bases de datos que nos proporcionan identidades para construir nuestros modelos, hacernos una idea de cuan bueno es nuestro método comparado con otros del estado del arte o incluso realizar comparaciones de nuestro escenario con otro escenario que posea características similares. Uno de estos escenarios son los relacionados con la diarización.

De forma usual, el término de diarización se refiere a un tipo de problema relacionado con técnicas focalizadas en la biometría de la voz, con lo que se consigue identificar/re-identificar a la persona mediante el habla. Siendo el objetivo de la diarización responder a la pregunta "¿Quién ha hablado en cada momento?". Pero, que sucede si intentamos responder a esta pregunta desde un punto de vista diferente, haciendo uso de técnicas basadas en visión por computador.

Por tanto, en esta tesis se indaga en los problemas de re-identificación de personas usando visión por computador en escenarios susceptibles de aplicar la diarización de personas. De forma metodológica, se revisan las técnicas de diarización tradicionales, así como el impacto que tienen los errores de diarización. Se demuestra que las estrategias tradicionales de diarización pueden no ser suficientemente precisas para determinadas aplicaciones. Por lo que proponemos en este trabajo diferentes metodologías para hacer frente a las carencias que afectan a las técnicas tradicionales.

"Un ganador es un
soñador que nunca
se rinde"

Nelson Mandela

Agradecimientos

Deseo agradecer la dedicación y esfuerzo que han tenido que realizar mis tutores, los Dres. Javier Lorenzo y Modesto Castrillón para la finalización de esta tesis doctoral. Pero, esto va más allá de una tesis, son un sinfín de situaciones que hemos vivido a lo largo de estos cuatro años. Para mí, aprender de los mejores en Canarias de este campo es un orgullo y un placer.

Durante el desarrollo de esta tesis han tenido lugar varias estancias de investigación. En la primera estancia fui acogido en el grupo de investigación de Robótica y Sistemas Autónomos de la Universidad del País Vasco, donde colaboré con los Dres. Itziar Irigoien y Basilio Sierra. Me integraron como un miembro más del grupo otorgándome una sensación de pertenencia a un grupo de personas maravillosas. Asimismo, me gustaría agradecer a mis tíos Pedro Marín y Rosi Balerdi por compartir momentos y tiempo conmigo. En mi segunda estancia, partí a Módena, una ciudad que no me llamó mucho la atención, pero poco a poco, gracias a los compañeros del AImageLab, donde tuve la suerte de colaborar, me hicieron cambiar mi forma de ver la ciudad. Agradecer a la Dra. Rita Cucchiara por brindarme la oportunidad de colaborar en un laboratorio tan dinámico. Ha sido un placer trabajar y compartir opiniones con los Dres. Simone Calderara y Andrea Palazzi. Como no iba a ser menos, me gustaría agradecer a otros miembros del laboratorio, como son Angelo Porrello, Stefano Pini o Luca Bergamini, por compartir también su tiempo de ocio. Al Dr. Federico Bolelli, espero que su proyecto personal siga hacia delante. Y en último lugar, a los Dres. Marcella Cornia y Lorenzo Baraldi que no solo teníamos con vernos en Europa, sino que coincidimos en el mismo congreso en América. Sin lugar a dudas, se los agradezco a todos ustedes como también a todas las personas que se cruzaron en mi vida en este periodo.

También me gustaría agradecer a los compañeros del laboratorio. David Castillo, con quien compartía mis pequeños hitos. Enrique Ramón, una gran persona con la que siempre se puede hablar y tomar un café. Además, al Dres. Mirko Marras y a Paola Barra, ambos realizaron una estancia de investigación, que hacen que investigar resulte divertido.

En último lugar, tengo que dar las gracias a mi familia, que ha sido y es la principal fuente de apoyo, siempre mostrando interés y positividad. También, a mi pareja, que siempre ha estado ahí apoyándome en todo momento con las decisiones que he tomado, aunque geográficamente la distancia nos separaba.

Grazie, mila esker, gracias, de otra manera nada de esto sería posible...

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

Resumen: En este capítulo se plantea la motivación por la cual se lleva a cabo esta tesis. Esta motivación da lugar a una serie de objetivos que han sido planteados conjuntamente entre el supervisor y el doctorando. Esta tesis doctoral y sus objetivos están avaladas por una serie de publicaciones que han sido sometidas a revisión por pares.

1.1. Motivación

En la actualidad existen infinidad de situaciones cotidianas que pueden ser automatizadas haciendo uso de sistemas inteligentes, desde el ajuste automático del termostato para una determinada persona hasta la detección de eventos anómalos en casa, como pueden ser caídas de personas dependientes o robos en el hogar. Para parte de estos problemas es imprescindible la detección e identificación de la persona usando una red de cámaras.

La monitorización de personas mediante redes de cámaras tanto en entornos de interior como de exterior, ha supuesto un incremento en la capacidad de adquisición de información visual que habitualmente se almacena durante el tiempo máximo que permite la legislación. El uso que se suele hacer de esta información es a posteriori, cuando se produce un evento, situación anómala o el etiquetado de las personas que aparecen en las grabaciones. Este análisis normalmente es realizado por un operador humano lo que supone una carga de trabajo bastante alta, que además es propensa a errores debido a lo monótono de la tarea.

Sin embargo, esta información visual almacenada puede servir a otros propósitos como es el análisis de las personas que acceden o se desplazan por el entorno cubierto por la red de cámaras [Wang, 2013, Lorenzo-Navarro et al., 2013, Zheng et al., 2015, Ristani et al., 2016]. Este análisis puede ir desde el simple conteo de las mismas hasta la obtención de rasgos biométricos como puede ser el sexo, la edad, descripción de la vestimenta o incluso parentesco. De esta forma emerge el concepto de Analítica de Personas como un símil con el de Analítica de Negocios en el entorno empresarial para el análisis de los datos. Un tipo específico del análisis es la diarización [Tranter and Reynolds, 2006, Barra-Chicote et al., 2011, Stolcke and Yoshioka, 2019, Tsipas et al., 2020, Ding et al., 2020], es decir, indicar quién y cuándo interviene en un coloquio o debate.

La diarización tiene lugar en múltiples escenarios, como pueden ser los programas de noticias, debates entre intervinientes, entrevistas, entre otros. Entre estos ámbitos, destaca por su significancia en la vida diaria de todos los ciudadanos de un país o comunidad, la actividad parlamentaria de sus representantes. En este sentido, disponer de información sobre las intervenciones de los mismos en las sedes parlamentarias es de gran importancia. Prueba de ello son las distintas leyes de transparencia que existen en la actualidad, para que los ciudadanos puedan fiscalizar la labor que hacen los representantes públicos. Por tanto, disponer de una herramienta que pueda etiquetar de forma automática las intervenciones de los parlamentarios en sede parlamentaria supondría un avance en el ámbito de la transparencia [Sánchez-Nielsen et al., 2017, Sánchez-Nielsen et al., 2019].

Sin embargo, para el desarrollo de esta herramienta, es necesaria la detección y caracterización de las personas. En este sentido, se han desarrollado diferentes técnicas de detección de personas [Viola et al., 2003, Dalal and Triggs, 2005, Enzweiler and Gavrila, 2008, Dollar et al., 2010, Kim et al., 2010, Xia et al., 2011, Dollar et al., 2011, Sermanet et al., 2013] como son los histogramas de gradientes orientados o clasificadores en cascada. Recientemente, se ha tendido por extraer las características haciendo uso del aprendizaje profundo [Zhong et al., 2018, Yu et al., 2020, Ye et al., 2021]. En el caso de la diarización, la detección de personas puede ir desde un problema sencillo cuando los planos y las condiciones de iluminación son constantes, a más complejo cuando la red de cámara proporciona diferentes poses y desde distintas ubicaciones de los intervinientes.

Una vez se ha detectado a la persona, es posible obtener una serie de descriptores basada en biometría blanda [Jain et al., 2004]. En aquellos escenarios donde se disponga de una vista frontal de la cara de la persona, a partir de ella se pueden obtener descriptores como el sexo [Moghaddam and Yang, 2002, Mäkinen and Raisamo, 2008, Castrillón-Santana et al., 2013, Galiyawala et al., 2018, Gonzalez-Sosa et al., 2018], y otros como la edad, etnia, etc. [Bourdev et al., 2011, Kumar et al., 2011, Dinca and Hancke, 2017].

Para escenarios donde la diferencia de tiempo entre las diferentes apariciones de una misma persona no es muy alta, la utilización de otras pistas como puede ser la ropa, incorpora una nueva fuente de información a la hora de desambiguar entre sujetos. Así, la descripción de vestimenta [Gallagher and Chen, 2008, Freire-Obregón et al., 2014, Lorenzo-Navarro et al., 2014] o la detección de diferentes estilos de ropa [Yang and Yu, 2011, Chen et al., 2012, Yamaguchi et al., 2012, Kalantidis et al., 2013, Ge et al., 2019, Zhang et al., 2020] ha empezado a captar la atención de la comunidad de visión por computador desde un punto de vista de las compras online, aunque se puede trasladar al objeto de esta tesis. Además, la información de la vestimenta junto a otra proporcionada por otros elementos ha sido explorada por parte de la comunidad científica para la determinación del sexo [Li et al., 2012, Liu et al., 2018, Lin et al., 2019].

1.2. Hipótesis de trabajo

Hoy por hoy, cada vez tienen más importancia los procesos de caracterización de personas que van introduciéndose en nuestra vida cotidiana. Tal como se plasma en la Motivación de esta tesis, ese aspecto de la cotidianidad que se va a tratar son principalmente las intervenciones políticas. Es por ello que nos preguntemos si *¿las metodologías actuales son efectivas en este ámbi-*

to?, ¿podrían mejorarse?. Esta hipótesis de la cual partimos, se puede dividir en varias sub-hipótesis que son las que van a dar forma a este documento.

- A la hora de analizar un vídeo, este dispone de una vasta cantidad de información visual que no se emplea, multitud de imágenes redundantes, de la que desconocemos si esa información hace que los resultados sean poco favorables. Por tanto, nos planteamos como primera sub-hipótesis si *¿es posible mejorar los resultados de re-identificación haciendo uso de la información esencial?*

- Estos escenarios son bastantes dinámicos en relación con los participantes, donde las personas intervienen y hacen alusión a otras desde múltiples localizaciones. Por consecuente, resulta complicado modelar de forma eficaz a las personas. Es por ello, que buscamos identificar si *¿el uso de las características locales para modelar una persona afecta la re-identificación de esta?*

- Existen características que afectan al proceso de re-identificación que son externas a los participantes. Los cambios de luminosidad, los tipos de cámaras que se usan, la distancia al objetivo, entre otras muchas. Esto nos suscita la siguiente cuestión, *¿afecta a la re-identificación el tipo de escenario en el que se está desarrollando?*

- Los algoritmos que se investigan actualmente son verificados en dispositivos que tienen altas prestaciones. Pero, *¿estas técnicas son capaces de exhibir un funcionamiento óptimo en dispositivos de bajas prestaciones?*

1.3. Objetivos

Partiendo de las hipótesis anteriores, se plantean los objetivos que se persiguen alcanzar en esta tesis doctoral:

- Estudio del estado del arte. Este es el objetivo básico en la realización de cualquier tesis doctoral ya que es fundamental conocer donde se encuentra la frontera del conocimiento en el campo de estudio para aportar nuevo conocimiento a partir de ese punto.

- Propuesta de un esquema de re-identificación para la detección de fotogramas claves en un entorno de grabación de vídeo para la diarización.

- Estudio de características locales de modelado de personas para facilitar su re-identificación en un problema de mundo abierto como es la diarización.

- Estudio de técnicas y su aplicación en la determinación del número de intervinientes en un debate o coloquio a partir de la información visual.
- Difusión de resultados. Este es otro de los objetivos básicos de cualquier investigación y es la difusión de los resultados obtenidos tanto en publicaciones científicas como en congresos tanto nacionales como internacionales.

1.4. Publicaciones y estructura de la tesis

Esta tesis doctoral está caracterizada por ser un documento extenso autocontenido, a diferencia de las tesis con formato de compendio de artículos. Sin embargo, este documento se enmarca bajo una serie de publicaciones en congresos y revistas de reconocido prestigio en el ámbito, dando lugar a priori a una justificación del trabajo desarrollado para esta tesis.

Esta tesis ha sido el resultado de 10 publicaciones en revistas y congresos que en su mayoría son de nivel internacional. Siendo 3 artículos en revistas del primer y segundo cuartil dentro del ámbito en la que se enmarca esta tesis [Marín-Reyes et al., 2019, Rodriguez et al., 2020, Freire-Obregón et al., 2021]. 6 publicaciones en congresos donde solamente una contribución de ellas fue presentada en un congreso nacional [Marín-Reyes et al., 2016, Lorenzo-Navarro et al., 2018, Marín-Reyes et al., 2017, Marín-Reyes et al., 2018, Marras et al., 2019a, Ortega-León et al., 2019], y además, destacar otra publicación en un *workshop* adscrito al Congreso CVPR (acrónimo en inglés de Computer Vision and Pattern Recognition), catalogado como un congreso A* bajo el ranking CORE (acrónimo en inglés de Computing Research and Education). Este congreso es uno de los más relevantes en visión por computador a nivel mundial. En último lugar, el comité organizador del ICPRAM (acrónimo en inglés de International Conference on Pattern Recognition Applications and Methods) del año 2019 seleccionó nuestra contribución como una de las 7 más relevantes de los 107 aceptados en el congreso, viéndose reflejado en un capítulo de libro [Marras et al., 2019b].

Además, me gustaría agradecer sinceramente a los coautores de estas publicaciones por su dedicación y esfuerzo desempeñado. En las siguientes subsecciones se muestran estas contribuciones en base al tipo de contribución: en revista, congreso o capítulo de libro.

1.4.1. Contribuciones en revistas JCR

- Marín-Reyes, P.A., Irigoien, I., Sierra, B., Lorenzo-Navarro, J., Castrillón-Santana, M., & Arenas, C. *ILRA: Novelty Detection in Face-*

Based Intervener Re-Identification. Symmetry, 11, 1154, 2019. Factor de impacto: 2.645 (JCR - Q2).

- Rodriguez, I., Zabala, U., Marín-Reyes, P.A., Jauregi, E., Lorenzo-Navarro, J., Lazkano E., & Castrillón-Santana, M. *Personal Guides: Heterogeneous Robots Sharing Personal Tours in Multi-Floor Environments*. Sensors, 20, 2480, 2020. Factor de impacto: 3.576 (JCR - Q1).

- Freire-Obregón, D., Rosales-Santana, K., Marín-Reyes, P.A., Penate-Sanchez, A., Lorenzo-Navarro, J., & Castrillón-Santana, M. *Improving User Verification in Human-Robot Interaction from Audio or Image Inputs through Sample Quality Assessment*. Pattern Recognition Letters, 149, 179, 2021. Factor de impacto: 3.255 (JCR - Q2).

1.4.2. Contribuciones en congresos

- Marín-Reyes, P. A., Lorenzo-Navarro, J., Castrillón-Santana, M., & Sánchez-Nielsen, E. *Shot classification and keyframe detection for vision based speakers diarization in parliamentary debates*. In Conference of the Spanish Association for Artificial Intelligence, Springer, pp. 48-57, 2016.

- Marín-Reyes, P. A., Lorenzo-Navarro, J., Castrillón-Santana, M., & Sánchez-Nielsen, E. *Who is Really Talking? A Visual-Based Speaker Diarization Strategy*. In International Conference on Computer Aided Systems Theory, Springer, pp. 322-329, 2017.

- Lorenzo-Navarro, J., Castrillón-Santana, M., Gómez, M., Herrera, A., & Marín-Reyes, P. A. *Automatic counting and classification of microplastic particles*. In ICPRAM 2018-Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods, 2018.

- Marin-Reyes, P. A., Palazzi, A., Bergamini, L., Calderara, S., Lorenzo-Navarro, J., & Cucchiara, R. *Unsupervised vehicle re-identification using triplet networks.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 166-171, 2018.

- Marras, M., Marın-Reyes, P. A., Lorenzo-Navarro, J., Castrillón-Santana, M., & Fenu, G. *AveRobot: An Audio-visual Dataset for People Re-identification and Verification in Human-Robot Interaction*. 2019.

- Ortega-León, C., Marín-Reyes, P. A., Lorenzo-Navarro, J., Castrillón-Santana, M., & Sánchez-Nielsen, E. *Video Categorisation Mimicking*

Text Mining. In International Work-Conference on Artificial Neural Networks, Springer, pp. 292-301, 2019.

1.4.3. Contribuciones en capítulos de libros

- Marras, M., Marín-Reyes, P.A., Lorenzo-Navarro, J., Castrillón-Santana, M., & Fenu, G. *Deep Multi-biometric Fusion for Audio-Visual User Re-Identification and Verification.* In: Pattern Recognition Applications and Methods. ICPRAM 2019. Lecture Notes in Computer Science, Springer, vol 11996, 2020.

Los siguientes capítulos de este documento están organizados de la siguiente forma, en primer lugar, en el Capítulo 2 se plasma la revisión bibliográfica sobre los aspectos que se han trabajado en esta tesis. En el Capítulo 3 se describe el problema a tratar. El Capítulo 4 versa sobre la metodología desempeñada, y por último, las conclusiones y líneas futuras de este trabajo se plasmarán en el Capítulo 5.

Capítulo 2

Revisión bibliográfica

Resumen: En este capítulo se detalla la bibliografía analizada para establecer un punto de partida del trabajo desarrollado, pero no solo como punto de partida, sino como línea de trabajo durante el desarrollo de esta tesis. En primer lugar, se analizarán los trabajos desarrollados en el ámbito de diarización, tanto visual como auditiva. Seguidamente, se plantean metodologías de detección de planos para incorporarlo en el proceso re-identificación de personas. Además, se estudia bibliografía relacionada con la detección de la novedad dentro del problema de re-identificación, como también, la problemática de la interacción de personas con máquinas en escenarios complejos. En último lugar, se propone bibliografía sobre otro tipo de problemas donde es fundamental hacer uso de sistemas inteligentes para la detección e identificación de objetos, de forma similar que se emplea para la detección de personas.

2.1. Diarización

(**a**) (**b**) (**c**)

Figura 2.1: Ejemplo de diferentes tipos de escenarios donde puede aplicarse la diarización. (**a**) Programas de noticias (® Televisión Canaria). (**b**) Debates entre personas (® RTVE). (**c**) Programas de entrevistas (® Antena 3).

La diarización de oradores (intervinientes) hace referencia a una problemática donde el objetivo primordial es dar respuesta a "¿quién y cuándo habla?". Este problema puede darse en diferentes tipos de escenarios como son los informativos, debates televisivos o programas de entrevistas entre otros varios (véase Figura 2.1). El audio de los participantes y vistas próximas a estos están disponibles para ser utilizados. Por lo que se podrían combinar ambas fuentes de información con el fin de crear un sistema más preciso [Anguera et al., 2012, El Khoury et al., 2014]. Adicionalmente, las vistas de los participantes son generalmente de primer plano donde la pose frontal es predominante, lo que permite la extracción de la información visual de la cara, en lugar de una apariencia general del participante.

En la actualidad se incrementa la popularidad del uso de técnicas duales, es decir, técnicas audio-visuales. En [Bredin and Gelly, 2016] se hace uso de las series de televisión para evaluar la técnica de diarización. Su propuesta está basada en la aplicación de agrupamiento sobre las imágenes de las caras de los participantes, para asignar el grupo de caras más representativa con el grupo de audio correspondiente. Este último se extrae del agrupamiento lineal del BIC (acrónimo en inglés de Bayesian Information Criterion) de la secuencia de audio. Finalmente, se usa un agrupamiento BIC regular para obtener la diarización final.

A diferencia de los autores anteriores, en [Gebru et al., 2017] se propone un sistema de detección de múltiples intervinientes que usa la posición de donde se originó la señal de audio. Otros autores, [Le et al., 2015], usan el sistema denominado LIUM para extraer la diarización del audio, donde aparecen subtítulos, y DPM (acrónimo en inglés de Deformable Part-based Model) para detectar la información visual de las caras. Un campo aleato-

rio condicional basado en seguimiento multiobjetivo es usado para seguir a los participantes. Después, se emplea una técnica de agrupamiento basado en distancias de similitud y medidas biométricas. Con el fin de asignar los nombres a estos grupos se calcula un etiquetado de intervinientes uno-a-uno para maximizar la duración de la concurrencia entre grupos y los nombres proporcionados por un reconocimiento óptico de caracteres.

En contraposición con los trabajos anteriores, en [Friedland et al., 2009], los autores no detectan las caras de las personas, en su lugar, detectan zonas de color piel usando los coeficientes de crominancia del tono de piel en el espacio de color YUV, donde posteriormente, se obtienen vectores de movimiento de estos bloques. Los MFCCs (acrónimo en inglés de Mel Frequency Cepstral Coefficients) de la fuente de audio son combinados con la representación visual extraída, usando la función de log-verosimilitud de las dos GMM (acrónimo del inglés de Gaussian Mixture Models).

Otras técnicas para anotar a los intervinientes están basadas en la detección de planos [Singh and Aggarwal, 2015], con lo que se consigue segmentar temporalmente el vídeo en múltiples fragmentos donde en cada uno solo aparecerá un único interviniente. Estos fragmentos se denotan como planos que a su vez contienen una secuencia de imágenes.

2.2. Detección de planos

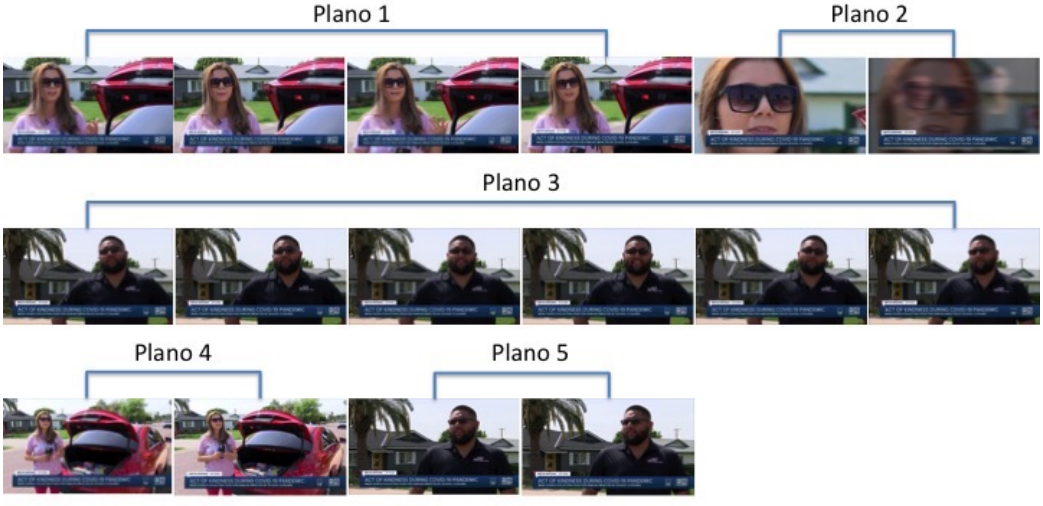


Figura 2.2: Secuencia de un fragmento de vídeo donde se señala el inicio y el fin de cada plano. Imágenes extraídas de Youtube (https://www.youtube.com).

La segmentación temporal de vídeos en planos que comparten las mismas características es importante ya que delimita el comienzo y el fin de la intervención de un participante en contextos donde aparece en primer plano una única persona. Por tanto, a la hora de identificar a una persona poseemos múltiples imágenes de la misma, por lo que se aumentan las probabilidades de acierto, ver Figura 2.2.

En [Sánchez-Nielsen et al., 2017] se hace uso de la segmentación de planos para asignar las diversas identidades a cada uno de los intervinientes que participan en las sesiones plenarias del Parlamento de Canarias. La detección de planos se basa en el cálculo de las diferencias de las distribuciones de color entre imágenes consecutivas, siendo un plano nuevo detectado cuando dichas diferencias exceden de un determinado umbral. Para ello se plantea un modelo basado en la distribución de color de la imagen en el espacio de color YCbCr donde se compararán fotogramas consecutivas haciendo uso de la divergencia de Kullback-Leibler.

Es evidente que es fundamental la reducción del tiempo de cómputo en la detección de planos. En [Gao and Ma, 2014] los autores critican que múltiples contribuciones se centran en la precisión de los métodos obviando el coste computacional. Proponen un algoritmo que es capaz de funcionar en tiempo real y de forma eficiente para aplicaciones de procesamiento de vídeo. Reduciendo el cálculo entre las imágenes, al tiempo que proporciona una precisión satisfactoria. Esta mejora de tiempo de cómputo se produce al reducir la región de interés, donde se obtiene el histograma de color y la información mutua con el fin de usarlos conjuntamente para medir la diferencia entre imágenes. Además, se utilizan para excluir falsos cambios de planos características locales, como la distribución de esquinas, de las imágenes.

En [Birinci and Kiranyaz, 2014] los autores proponen una novedosa técnica para realizar detecciones de planos de forma rápida y precisa. Hacen uso de reglas perceptivas humanas y de conceptos basados en la búsqueda de información [Shneiderman, 1996]. Los vídeos redundantes no son procesados con lo que consiguen reducir el coste computacional. Para detectar los planos hacen uso de características locales en las imágenes, dando lugar a un nuevo plano ante la existencia de discontinuidades en estas características locales. Un factor importante de esta técnica es que es totalmente genérica, puede ser aplicada a cualquier tipo de vídeo sin necesidad de realizar una fase de entrenamiento o ajuste de parámetros al algoritmo.

No solamente es importante identificar el plano sino escoger la imagen que mejor represente a este. En [Baraldi et al., 2016] plantean un sistema de detección de planos y escenas, como la selección de la imagen clave en vídeos. Las escenas agrupan una colección de planos consecutivos que a su vez será representado por una única imagen por plano, siendo la más representativa

por plano. La imagen clave se selecciona por la mediana de la distribución de imágenes por plano.

Estas técnicas expuestas como bien se ha comentado anteriormente pueden ser empleadas en los problemas de diarización. Asimismo, puede considerarse un problema similar al de re-identificación de personas. En los sistemas de diarización hay que identificar a la persona que está hablando. La diferencia recae en los escenarios que se emplean. La re-identificación de personas es mayoritariamente considerada en escenarios de videovigilancia, donde no hay audio, y las vistas de las personas son alejadas.

2.3. Re-identificación de personas

Figura 2.3: Ejemplo de un problema de re-identificación de personas. *Imágenes extraídas de Race Photos (https://www.racephotos.es/).*

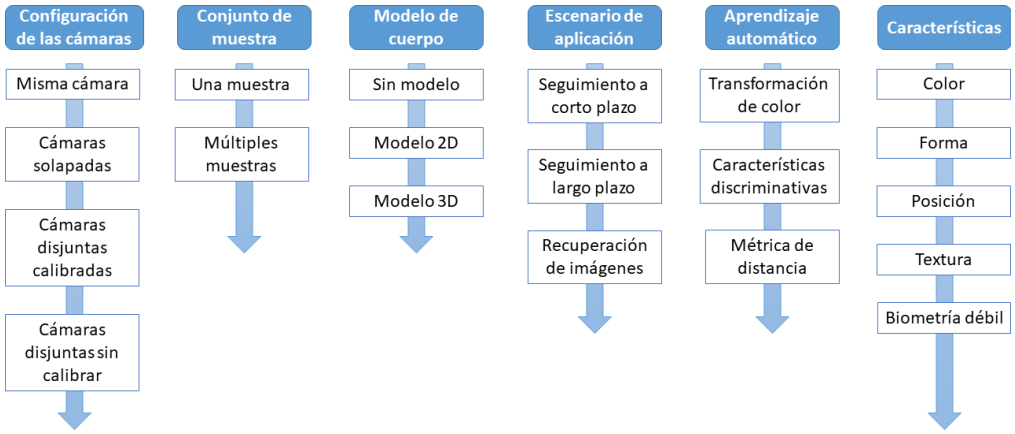
La re-identificación de personas es el proceso de reconocer a una persona

entre otras cámaras donde las grabaciones generalmente no están solapadas [Gheissari et al., 2006, Vezzani et al., 2013, Prosser et al., 2010, Roth et al., 2014, Bedagkar-Gala and Shah, 2014, Liao et al., 2015]. Comúnmente, se usa el término del inglés *probe* para referirse a la imagen del individuo que tiene que ser identificado, y el término en inglés *gallery* para el conjunto de imágenes de personas conocidas por el sistema donde el probe tiene que ser identificado, ver Figura 2.3. Los problemas de re-identificación pueden clasificarse en diferentes categorías atendiendo de la dimensión que se considere [Vezzani et al., 2013]: conjunto de muestra, modelo de cuerpo, escenario de aplicación, aprendizaje automático, configuración de las cámaras y características; ver Figura 2.4.a para más detalle. En [Bedagkar-Gala and Shah, 2014] proponen otra taxonomía donde inciden en la presencia obligatoria o no del probe en el gallery para cada una de las dimensiones de esta, ver Figura 2.4.b. Denominando mundo cerrado o conjunto cerrado al escenario que es similar al problema clásico de re-identificación de personas, donde cada probe tiene que existir en el gallery. Por el contrario, los autores denominan mundo abierto o conjunto abierto cuando el probe necesariamente no pertenece al gallery, el cual evoluciona dinámicamente, agregando nuevas identidades a medida que tiene lugar el proceso de re-identificación.

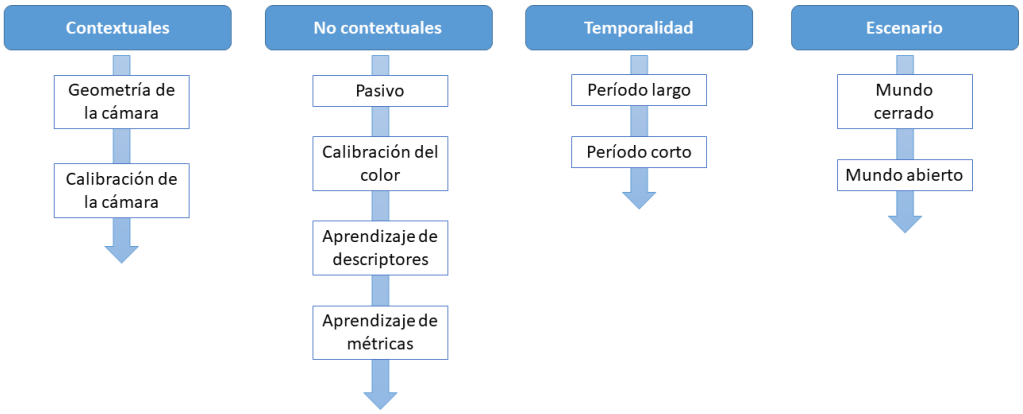
En los escenarios de re-identificación de mundo cerrado basados en características (Figura 2.4), en [Bazzani et al., 2013] Bazzani et al. proponen dividir el cuerpo de la persona en múltiples porciones. En estas se extraen las características usando histogramas ponderados sobre el espacio de color HSV. Otras características se extraen mediante un agrupamiento aglomerativo de los píxeles de la imagen y el cálculo de parches de textura.

Además, en la actualidad, se ha puesto en valor el aprendizaje de métricas de distancia (*metric learning*) para los problemas de re-identificación de conjunto cerrado, Figura 2.4. El objetivo de estas técnicas es proyectar la representación de los individuos en un espacio de características donde las muestras del mismo individuo están más cerca y las de diferentes individuos están más separadas. En [Tao et al., 2016] los autores proponen el aprendizaje KISS (acrónimo del inglés de Keep It Simple and Straightforward), mejorando el aprendizaje de métricas utilizando una regularización para suprimir el efecto de valores propios más grandes en las matrices de covarianza. Asimismo, en [Yu et al., 2017], los autores describen una técnica para encontrar el espacio común entre diferentes vistas de cámaras en un contexto no supervisado. Así, un K-Medias se usa para agrupar las imágenes de las personas de las diferentes vistas.

Las redes neuronales son también generalmente usadas para proyectar las muestras en un nuevo espacio de muestras. En este sentido, en [Ustinova et al., 2017] los autores dividen la imagen en tres regiones y usan estas como



(**a**) Dimensiones propuestas por Vezzani et al.



(**b**) Dimensiones propuestas por Bedagkar-Gala y Shah.

Figura 2.4: Ejemplos de dimensiones propuestas que pueden afectar a los problemas de re-identificación.

entrada a una red bilineal para agregar en un vector de características. Estos vectores se usan para obtener el nuevo espacio de características embebidas usando una red siamesa. Esta arquitectura es generalmente empleada para verificar las muestras de entrada de la red. En [Zheng et al., 2017], los autores también añaden una fase de identificación de personas al modelo.

Hay que tener en cuenta que en las propuestas presentadas se asume que el probe pertenece al gallery, lo que lo hace un problema menos complejo de lo que en realidad puede llegar a ser. Por lo que no refleja en su totalidad la problemática de los escenarios de videovigilancia, donde la persona no tiene que pasar obligatoriamente por todas las zonas donde las cámaras están distribuidas, sino que solamente podría pasar por una única zona donde hay

una cámara. Es por ello, que los problemas de mundo abierto dan mejor solución a esta particularidad.

2.4. Detección de la novedad en personas

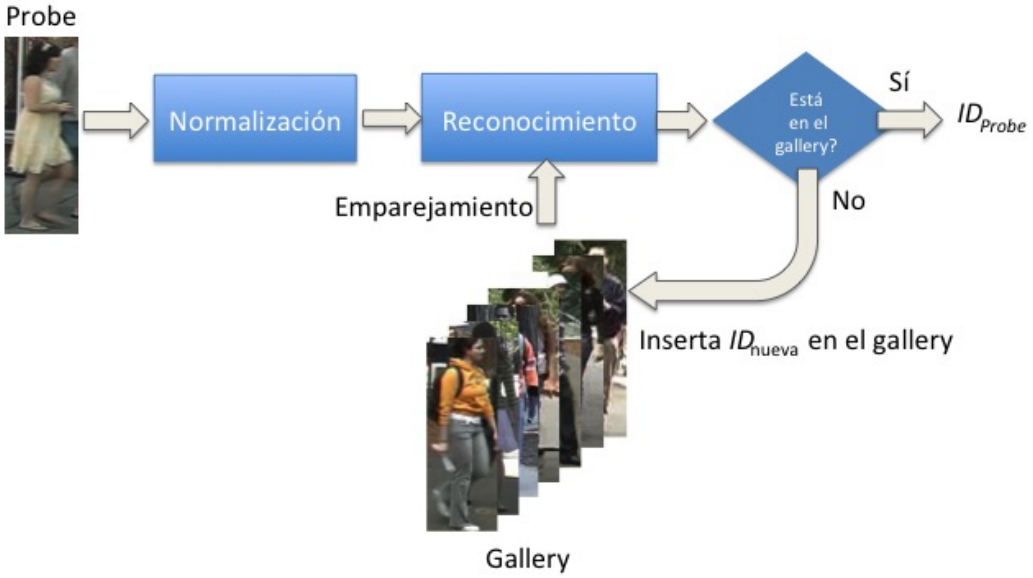


Figura 2.5: Esquema generalizado de los problemas de detección de novedad.

En los escenarios de re-identificación de mundo abierto, en primer lugar, es necesario decidir si el probe pertenece al gallery o no. Si el probe pertenece a este, un proceso de emparejamiento es llevado a cabo; si no, el probe es añadido al gallery como una nueva identidad. La primera etapa en los escenarios de re-identificación de mundo abierto es muy parecida a los problemas de detección de la novedad [Markou and Singh, 2003, Chandola et al., 2009, Pimentel et al., 2014], que se refiere a la identificación de individuos nuevos o desconocidos, que no estaban previamente registrados en el sistema. Esos individuos se denominan atípicos en oposición a los registrados, a los que se hace referencia como típicos.

Como se mencionó anteriormente, los escenarios problemáticos actualmente en el campo de re-identificación son aquellos relacionados con problemas de mundo abierto, donde la detección de novedad es imprescindible (Figura 2.5). La detección de novedad se utiliza en un gran tipo de contextos, como [Yong et al., 2012] en escenas de la vida silvestre; [Clifton et al., 2013] para series temporales de signos vitales en cirugía de cáncer gastrointestinal. Asimismo, el diagnóstico de enfermedades dérmicas y el análisis del cáncer

linfático han sido tratados en [Irigoien and Arenas, 2008] o en problemas relacionados con la robótica [Boucenna et al., 2016].

Relacionado con la re-identificación de personas, pero utilizando señales de audio, en [Markov and Nakamura, 2008] los autores proponen usar la detección de novedad en un sistema de diarización de intervinientes. Aplican un umbral basado en probabilidad, dependiendo del sexo del hablante; y se normaliza usando la media y la desviación estándar. Este umbral determina los intervinientes típicos/atípicos.

Sin embargo, esta tesis se centra en problemas de re-identificación basados prioritariamente en la información visual. En [Zheng et al., 2012] los autores proponen un novedoso método de *transfer ranking* para dos tipos de verificación, la verificación de múltiples muestras y la de una muestra, ver Figura 2.4(a), en un problema de *bipartite ranking*. Aplican RankSVM y la comparación de distancia relativa probabilística para obtener el modelo, que optimiza un parámetro marginal basado en las variaciones típicas en la clase y entre clases, y las variaciones entre clases entre imágenes típicas y atípicas.

Asimismo, en [Chan-Lang et al., 2017] los autores presentan una técnica de aprendizaje supervisado del subespacio donde el objetivo es aprender una transformación lineal de las características mediante la optimización de una función de coste relacionada con la proporción de pares clasificados erróneos de los positivos y negativos. Para determinar la presencia de un probe en el gallery introducen un parámetro de margen tal que los pares cuya distancia es inferior al umbral se consideran pertenecientes al gallery y no pertenecientes a esta en el caso contrario.

En [Zhu et al., 2018] los autores introducen una nueva configuración de búsqueda de re-identificación de personas donde las características principales son: una amplia población de búsqueda de probes, búsqueda rápida de vista disjunta y escaso número de identidades de personas en la fase de entrenamiento. Sobre esta configuración, obtienen un conjunto de características de la correlación de la identidad en las vistas cruzadas y la verificación de la discriminación de la identidad. De la misma manera que los autores anteriores, la detección de novedad se basa en un umbral sobre la distancia entre representaciones de los individuos.

En los últimos años, los problemas de re-identificación de mundo abierto se han tratado con el aprendizaje profundo, en particular se utilizan redes generativas. Por ejemplo, en [Deng et al., 2018] se presenta un método no supervisado de adaptación al dominio que genera muestras para un aprendizaje efectivo en el dominio objetivo. Esto se realiza bajo el supuesto de que los conjuntos de datos en diferentes dominios de re-identificación tienen conjuntos de identidades completamente diferentes. Por lo tanto, una imagen transferida debe tener una identidad diferente de cualquier imagen objetivo.

De esta manera, se utiliza una Red Adversaria Generativa de Ciclo (CycleGAN) [Zhu et al., 2017] para transferir imágenes de un dominio fuente a un dominio objetivo. Luego, se usa una red siamesa para generar un nuevo espacio de muestras donde las imágenes de diferentes personas están alejadas y las imágenes de la misma persona son próximas, con el objetivo de clasificar una muestra como típica o atípica.

Además, en [Li et al., 2018] los autores aprovechan el beneficio de integrar imágenes de personas generadas. Por un lado, usan un discriminador de personas para verificar si la imagen generada es una persona o no. Por otro lado, un discriminador objetivo identifica si una persona pertenece al conjunto de datos o no. El vector de características se extrae de la última capa completamente conectada del discriminador objetivo y se usa un umbral para determinar la novedad de la persona.

Las imágenes son adquiridas comúnmente con el mismo tipo de sensor para cada una de las técnicas mencionadas en esta sección. Un problema más complejo es el uso de diferentes tipos de sensores en la identificación de una persona, como es el caso de los escenarios donde interactúan varios robots con las personas. Estos poseen diferentes sensores, lo que hace que las imágenes se vean afectadas por la diferente respuesta de los sensores a la iluminación o a diferentes posiciones del sensor.

2.5. Interacción hombre-máquina

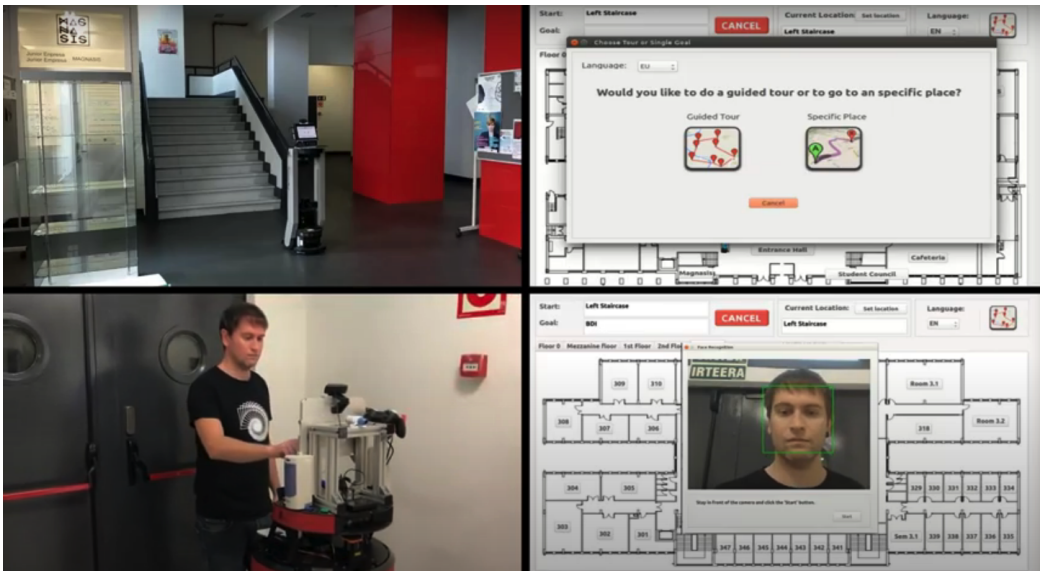


Figura 2.6: Sistema de robots asistentes colaborativos.

En el campo del reconocimiento biométrico en interacción hombre-máquina, HRI (acrónimo en inglés de Human Robot Interaction), ha estado dominado por técnicas que integran características diseñadas ad-hoc relacionadas tanto con la biometría dura, principalmente facial, y la blanda, como el sexo, la edad o la altura [Cielniak and Duckett, 2003, Cruz et al., 2008]. La combinación de características biométricas audio-visuales fueron desarrolladas por [Martinson and Lawson, 2011] donde realizaban reconocimiento facial empleando redes neuronales básicas y reconocimiento de intervinientes con GMMs (acrónimo del inglés de Gaussian Mixture Models). Para incrementar la robustez en escenarios no controlados, [Ouellet et al., 2014] combinan la identificación de voz y cara con características biométricas.

En [Correa et al., 2012] los autores emplean caras modeladas en los espectros térmicos y visuales para el mismo objetivo. En caso de mala iluminación, su enfoque se basa en la información térmica, mientras que la información térmica y visual se complementan entre sí en buenos escenarios de iluminación. Las características independientes de la iluminación también fueron utilizadas por [Koide and Miura, 2016].

A diferencia de los autores anteriores, en [Sinha et al., 2013] se usaron puntos de referencia corporal, con lo que se estimó el esqueleto de la persona, para detectar la forma de andar de la persona y calcular las características de esta. La selección y clasificación de características se realizó con redes neuronales adaptativas. Para gestionar las limitaciones de las características de esqueleto y color al tratar con oclusiones y orientación, [Cosar et al., 2017] y [Liu et al., 2017] presentaron enfoques basados en RGB-D utilizando características del volumen corporal.

Otro escenario diferente donde existe una interacción entre hombre y robot es en las plataformas de aprendizaje online. El 14 de marzo de 2020 se inicia el estado de alarma en todo el territorio español debido al avance de la pandemia de COVID-19 [Shaoshuai et al., 2020], haciendo peligrar la educación en España. En el sistema educativo se impone por necesidad un cambio de modelo de enseñanza, pasamos de una enseñanza presencial a una enseñanza online. En este sentido, es importante tener los recursos necesarios para que el alumnado pueda desarrollar las diferentes sesiones con éxito. Además, la evaluación del alumnado tiene que ser llevada a cabo con todo el rigor posible por lo que es necesario avanzar en el control del alumnado a la hora de asistir a las clases telemáticas, entrega de trabajos o realización de exámenes en línea. Es por ello que se hace uso de técnicas de biometría para la identificación de estudiantes que están desarrollando un trabajo detrás de la pantalla. En [Fenu et al., 2018] se desarrolla un sistema transparente y continuo de autenticación de estudiantes en plataformas online. Este es un sistema multibiométrico independiente del dispositivo que realiza una fusión

de la puntuación obtenida para las diferentes modalidades biométricas (cara, voz, tacto, ratón, pulsación de tecla) en función del dispositivo utilizado y la actividad que se lleve a cabo con este. Asimismo, en [Fenu and Marras, 2018] los autores mejoran el sistema de autenticación haciendo uso de características biométricas faciales y de comportamiento (movimientos táctiles y manuales).

El reconocimiento de las personas ha tenido poca atención en aplicaciones reales de re-identificación facial en robots sociales, como se menciona en [Wang et al., 2018b] pero es un problema que ha recibido la atención de los investigadores de Visión por Computador. Uno de los primeros intentos en este ámbito fue probablemente la investigación llevada a cabo por [Wong et al., 1995] donde el problema de reconocimiento de personas es dividido en diferentes pasos y usaban un método no supervisado de extracción de características. Estas eran usadas como entradas a una red neuronal que se encarga de distinguir las caras de las personas y asignar una identidad a estas. El sistema fue desplegado en un robot, Cybermotion K2A, con una tasa de exactitud del 70 %.

Múltiples son las características que se pueden emplear para identificar personas. En [Martinson et al., 2013] se comparan tres rasgos de biometría blanda (vestimenta, complexión y altura) para ser usados por un robot humanoide en un entorno social donde tiene que identificar a las personas. Asimismo, se pueden utilizar otras entradas sensoriales como la voz y el tacto, pero el reconocimiento facial basado en visión es uno de los métodos más seguros para reconocer a las personas [Sinha et al., 2006].

Capítulo 3

Descripción del problema

Resumen: En este capítulo se describirá el problema al que se hace frente en esta tesis. Este problema consta de varios procesos que serán especificados. Además, se detallarán los escenarios donde se fundamenta nuestra propuesta de trabajo.

3.1. Fases en el proceso de re-identificación

En el proceso de re-identificación desarrollado, se busca resolver las cuestiones planteadas en nuestras hipótesis de trabajo, ver Sección 1.2. Es por ello, que se pretende realizar un análisis exhaustivo de las diferentes fases involucradas en el proceso de re-identificación. Asimismo, una característica relevante que influencia el desarrollo de este trabajo es la aplicación o el desarrollo de modelos que sean aplicables en escenarios de re-identificación de mundo abierto. Por tanto, en el Capítulo 4 se plasmarán estas cuestiones en aras de ir analizándolas.

Como se ha descrito en el capítulo anterior, existen múltiples casuísticas que hacen que el problema de re-identificación sea diverso. No existe una única forma de afrontar un problema de re-identificación particular. En este trabajo nos centramos en los siguientes aspectos:

- Re-identificación de personas en debates
- Detección de la novedad en re-identificación
- Interacción hombre-máquina en robots asistentes
- Diseño de base de datos en interacción hombre-máquina
- Re-identificación multimodal

En primer lugar, en los debates, es importante la extracción de los fotogramas claves a la hora de aplicar la re-identificación porque se evita el uso de imágenes redundantes o que aportan poco valor al proceso que nos concierne. Por tanto, se buscan imágenes donde aparezca una persona, que no esté muy alejada y que tenga una pose que favorezca la re-identificación.

En segundo lugar, hemos reseñado que vamos a tratar problemas de mundo abierto, por lo que es necesario emplear una heurística que sea capaz de determinar que persona es novedosa para el sistema. Por un lado, el sistema parte de cero para desarrollar su base de datos, por tanto, tendrá que añadir de forma incremental las personas con sus identificadores correspondientes. Por el otro lado, tiene que ser capaz de verificar si dos personas son la misma o diferente persona.

En tercer lugar, se busca que el trabajo desarrollado tenga una utilidad práctica. Para ello, se utilizan los conocimientos adquiridos en este trabajo para otorgar a varios robots la capacidad de decidir si una persona es la misma. La utilidad de estos robots se emplearán en el guiado de personas en un edificio.

En último lugar, es de interés comprobar que el trabajo desarrollado puede emplearse en escenarios complejos. Para realizar este análisis, se disponen de varias cámaras con diferentes características para grabar un vídeo en varias localizaciones dentro de un edificio. Estas localizaciones exhiben características particulares, como puede ser que una zona sea más luminosa que otra, la distancia de la persona o la altura desde donde se adquiere la imagen. Además, se van a emplear diferentes características biométricas para caracterizar al interviniente, no solo se va a emplear imágenes, sino que también se va a emplear audio.

Además, las metodologías propuestas se van a desarrollar en diferentes escenarios. Esta variedad va a permitir que los resultados sean generalizables con el tipo de escenario que tratamos.

3.2. Escenarios de aplicación

A la hora de aplicar visión por computador, es crucial realizar un análisis del entorno donde se va a desarrollar. Qué situaciones afectan a estos escenarios y el cómo se va a hacer frente a estos. Con este fin, se hace uso de tres escenarios: El Parlamento de Canarias, las instalaciones de la Escuela Universitaria de Informática de la Universidad de Las Palmas de Gran Canaria (ULPGC) y de la Facultad de Informática de la Universidad del País Vasco (UPV) de San Sebastián (Figura 3.1). A continuación se detalla cada uno de estos escenarios:

- Parlamento de Canarias (1ª fila de la Figura 3.1): es un escenario que recoge las intervenciones de los políticos en las sesiones parlamentarias del Gobierno de Canarias. El objetivo es re-identificar a los intervinientes que participan en estas sesiones. En este escenario se dispone de tomas desde diferentes puntos para los intervinientes, variando los ejes o aplicando zoom. Además, suelen haber variaciones en las condiciones lumínicas como en ajustes automáticos de color o cambios de vistas. Hay multitud de personas y se pueden producir oclusiones con respecto a la persona que esta interviniendo en la sesión parlamentaria. En estas sesiones, intervienen entre 2 a 33 parlamentarios.

- Instalaciones de la Escuela Universitaria de Informática de la ULPGC (2ª fila de la Figura 3.1): tiene lugar en un escenario de interior típico de robots de asistencia. Por tanto, se recopilaron vídeos de diferentes personas interactuando ante las cámaras con el fin de simular que se solicita la ayuda a un robot. Teóricamente, estos robots están distribuidos en diferentes plantas, el edificio de la Escuela de Ingeniería Informática

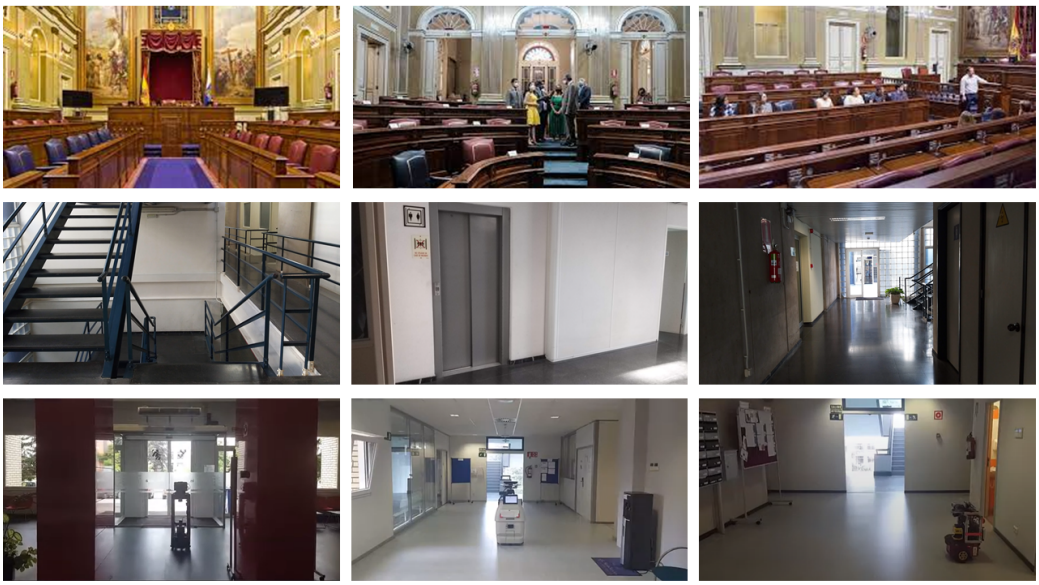


Figura 3.1: Imágenes sobre los diferentes escenarios en los que se desenvuelve este trabajo.

cuenta con tres plantas, donde se usaron ocho cámaras para captar las interacciones. Además, las cámaras estaban situadas a alturas diferentes. Se situaron dos en la primera planta, tres en la segunda y tres en la tercera; con las que se grababan tres localizaciones adicionalmente. Con esto se consiguió establecer un escenario complejo donde intervinieron 111 personas grabadas por diferentes cámaras que plantean cambios en la iluminación, geometría y resolución de las imágenes y variaciones en sonido. Además, el audio tenía ruido de fondo, donde se perciben risas, conversaciones y la acústica que genera el lugar.

- Instalaciones de la Facultad de Informática de la UPV de San Sebastián (3ª fila de la Figura 3.1): tiene lugar en un escenario de interior, se emplean robots para adquirir las imágenes. Para ello, se ha desplegado un módulo de visión por computador para la interacción entre robots colaborativos con el objetivo de guiar a 56 usuarios entre las diferentes plantas del edificio. Son robots heterogéneos, no pueden cambiar de plantas por las características de estos. Los usuarios son visitantes casuales, por lo que no se almacena la información en el sistema, por tanto, las personas son desconocidas para el sistema. Este escenario está constituido por un edificio de cuatro plantas que cuentan con dos escaleras laterales y un solo ascensor que permite a las personas moverse entre estos.

Capítulo 4

Metodología

Resumen: En este capítulo se detalla el desarrollo de los principales puntos de esta tesis doctoral. La Sección *Re-identificación de personas en debates* se centra en la detección de personas y el preprocesado de estas para mejorar el tiempo de cómputo y hacer más fiable el sistema. En la Sección *Detección de la novedad en re-identificación*, se expone una técnica novedosa para el reconocimiento de personas en un contexto de mundo abierto donde es necesario determinar si una persona ha sido registrada o no por el sistema. En la siguiente Sección, *Interacción hombre-máquina en robots asistentes*, se pone en explotación los métodos de visión por computador analizados previamente en múltiples robots de servicios. Dadas las dificultades encontradas a la hora de seleccionar bases de datos relacionadas con esta temática, en la Sección *Diseño de base de datos en interacción hombre-máquina* se propone un conjunto de datos desafiante dentro de la problemática de HRI donde se posibilita la aplicación de algoritmos multimodales. En último lugar, en la Sección *Re-identificación multimodal* se hace uso del conjunto de datos propuesto en la sección anterior usando metodologías multimodales.

4.1. Re-identificación de personas en debates

Figura 4.1: Diferentes vistas de las cámaras durante una sesión parlamentaria.

4.1.1. Escenario

Las sesiones parlamentarias son un escenario desafiante dado que las grabaciones no proporcionan un campo único de visión centrado en cada intervención. En lugar de esto, una red de cámaras adquieren diferentes vistas del parlamento. Además, las especificaciones de estas son heterogéneas y existen cambios tanto en los ejes horizontal y vertical, como la realización de zoom. Este escenario está caracterizado también por la vestimenta similar entre los intervinientes, cambios de condiciones lumínicas y los ajustes automáticos de color durante la grabación, variaciones de las vistas entre cámaras cuando el interviniente imparte un discurso, fondos abarrotados y oclusiones. Dado que nuestro propósito es etiquetar de forma automática cada interviniente durante su discurso parlamentario, el sistema desarrollado debe detectar los diferentes tipos de planos en aras de procesar solo los planos válidos, evitando cálculos redundantes.

La literatura reciente sobre visión por computador es rica en técnicas de detección de personas [Nguyen et al., 2016]. Existen diferentes patrones visuales que han sido tenido en cuenta para ese propósito: el rostro/cabeza, cuerpo superior, cuerpo completo o solo las piernas. Para este escenario, incluso si el interviniente estuviera mirando a la audiencia en lugar de a la cámara, su pose será típicamente frontal. Por lo tanto, el interviniente podría estar de pie rodeado por la audiencia mientras ellos están sentados cerca de él/ella. Por lo tanto, los detectores de cara y parte superior del cuerpo se ajustan a las restricciones del problema según el tipo de plano (Figura. 4.1).

El sistema desarrollado esta formado por seis módulos, ver Figura 4.2. La imagen de entrada es procesada por un detector de planos, el cual determina si la imagen pertenece a un plano nuevo para su posterior clasificación entre varios tipos de planos. La imagen se procesa por un detector de cuerpo superior si el plano satisface unas condiciones determinadas, generando una imagen recortada de la región de interés donde se aplica un detector de

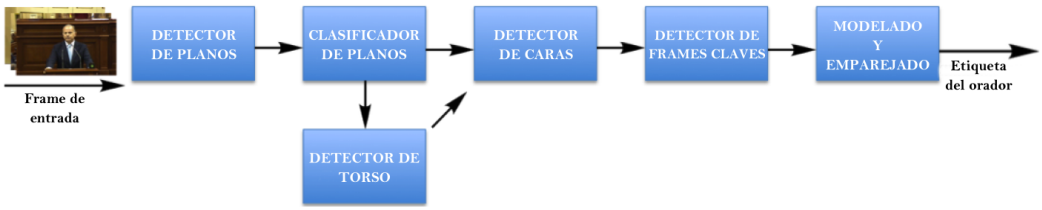


Figura 4.2: Descripción general del sistema.

caras para determinar el área del rostro del interviniente, como la posición de los ojos y la boca. Estas posiciones numéricas se emplean en el cálculo de un coeficiente para verificar que la cara corresponde a una cara real y con buenas características. Esta área de la cara se modela por características visuales y emparejado con el interviniente con mayor similitud, o en el caso que no fuera similar a ningún interviniente, se crearía una nueva identidad. A continuación se detallan cada uno de estos módulos.

4.1.2. Desarrollo

4.1.2.1. Detección de planos

La detección de planos es necesaria en los procesos de indexación automática de vídeos. Su detección proporciona semántica sobre el flujo de vídeo procesado. Comúnmente se emplean técnicas basadas en el cálculo estadístico y la definición de un umbral entre imágenes. En este apartado en particular, se usa una técnica de umbralizado basada en la comparación de fotogramas consecutivos usando la divergencia de Kullback-Leibler (KL) [Cover, 1999] entre los histogramas del espacio de color HSV [Sánchez-Nielsen et al., 2017].

4.1.2.2. Clasificador de planos

Figura 4.3: Tipos de planos. (**a**) Primer plano, (**b**) Plano medio, (**c**) Plano largo, (**d**) Otros.

En los debates parlamentarios, los diputados pueden participar desde sus asientos o desde el estrado. Por ello, denominamos planos próximos a aquellos

donde solo los intervinientes aparecen y la pose de la cara es principalmente frontal. El segundo tipo de planos lo nombramos plano medio, donde el interviniente es el sujeto principal de la imagen pero suele aparecer rodeado por otros diputados. El tercer tipo de planos corresponden a planos largos, las vistas son generales del parlamento. El último tipo de planos lo forman encuestas y títulos. En la Figura 4.3 se pueden distinguir estos tipos de planos.

Una CNN (acrónimo en inglés de convolutional neural network) [LeCun et al., 1998] ha sido empleada con el fin de diferenciar estos cuatro tipos de planos. La entrada a la red tiene una dimensión de $227 \times 227 \times 3$, imágenes en RGB. Tres capas de convolución seguidas de una función de activación ReLU, seguido de un *pooling* máximo y una capa de normalización. Las siguientes dos capas corresponden a capas neuronales totalmente conectadas, ambas con un tamaño de 512 y ReLU como función activadora. La última capa corresponde a la capa de clasificación formada por 4 salidas y softmax como función de activación.

4.1.2.3. Detección del interviniente

Figura 4.4: Imágenes de detección de caras utilizando detectores de rostro y parte superior del cuerpo.

Como se ha mencionado anteriormente, la configuración del escenario nos brinda la posibilidad de introducir ciertas restricciones para identificar el tipo de plano al que pertenece el interviniente. Los planos largos son excluidos porque en general no proporcionan información visual válida para procesar, las áreas de interés son muy pequeñas. Un detector de caras es usado en primeros planos debido a que estos planos son próximos al interviniente y solo aparece este. Los planos medios tienen una limitación a la hora de detectar las caras porque el detector localiza todas las caras en la imagen. Aplicar un criterio de selección a partir del mayor tamaño de la cara no funciona correctamente debido al ángulo de la cara o la morfología de esta. En la fila

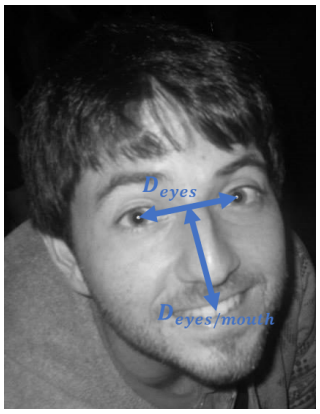


Figura 4.5: Imagen de ejemplo donde se muestran los puntos de los cuales se obtienen las distancias para el cálculo de ratio.

superior de la Figura 4.4 se muestran algunos ejemplos de este problema. Es importante tener en consideración que en este tipo de planos el interviniente está de pie. Esta situación puede ser detectada por un detector de cuerpo superior [Castrillón et al., 2011], para la posterior detección de la cara del interviniente. En particular, se han usado los detectores basados en la técnica de Viola-Jones [Viola and Jones, 2004]. En la fila inferior de la Figura 4.4 se muestran las detecciones correctas de la cara del interviniente en color verde, en azul la detección del cuerpo superior.

4.1.2.4. Extracción de fotogramas claves

Un fotograma clave es la imagen que define la representación relevante de un plano. La obtención de estos fotogramas claves reducen la cantidad de imágenes que el sistema tiene que modelar y además, se suprimen posibles errores acarreados en la fase de detección del interviniente. Hay diferentes métodos para extraer fotogramas claves [Sujatha and Mudenagudi, 2011]. Proponemos el uso de la detección de fotogramas claves usando las interdistancias entre los elementos faciales, distancia de los ojos y distancia entre el punto medio de los ojos y la boca. Se analizó estadísticamente la influencia de estas medidas para determinar un coeficiente, ver Ecuación 4.1, el cual representa una medida adimensional.

$$c = \frac{D_{eyes}}{D_{eyes/mouth}} \tag{4.1}$$

donde c representa el ratio entre la distancia de los ojos (D_{eyes}) y la distancia entre el punto medio de los ojos y la boca ($D_{eyes/mouth}$), ver Figura 4.5. En la

Ecuación 4.2 se plantea las reglas de decisión en base al análisis estadístico desarrollado entre los vídeos.

$$frame_i \text{ is keyframe} \begin{cases} \text{True} & \text{if } shotType \text{ is } Próximo \text{ and } 0{,}76 \leq c \leq 0{,}82 \\ \text{True} & \text{if } shotType \text{ is } Medio \text{ and } 0{,}83 \leq c \leq 0{,}89 \\ \text{False} & \text{Otherwise} \end{cases} \tag{4.2}$$

donde $frame_i$ corresponde a cada imagen del vídeo y $shotType$ referencia al tipo de plano del fotograma.

4.1.2.5. Modelado del interviniente y emparejado

Una vez la cara del interviniente se detecta, se toman tres regiones de interés en el modelado. El área del rostro se procesa haciendo uso de Histogramas de Gradientes Orientados (HOG) aplicando una rejilla de 3×3 con la que conseguimos nueve celdas de HOG que son concatenadas. Otra área de interés es la que abarca la parte superior de la cabeza, donde se puede obtener información del peinado e información grosera sobre la cara, ya que se hace uso de una rejilla de 2×2 para calcular el HOG en estos. Por último, el color de la vestimenta se usa para modelar al interviniente. Se emplea el histograma extraído del espacio de color YCbCr de la región que está por debajo de la cara. Estos tres tipos de descriptores son los que caracterizan a la persona.

El proceso de emparejado entre intervinientes se lleva a cabo haciendo uso de dos medidas de similitud. Se comparan los descriptores HOG empleando la distancia Coseno y para comparar el histograma de color se emplea la divergencia de KL. Para cada plano donde aparece un interviniente se comparan las imágenes modeladas con las imágenes previamente procesadas de los planos anteriores, asignando la identidad del interviniente al que menor distancia proporcione. En el caso de que la distancia supere un determinado umbral, se considera una identidad nueva.

4.1.3. Experimentos y resultados

Para evaluar experimentalmente este sistema se han utilizado vídeos del Parlamento de Canarias[1]. Consisten en 31 vídeos, que además, cuentan con la diarización manual de los mismos que sirve como anotación para verificar esta metodología. En la Tabla 4.1 se muestran en la parte izquierda las características de los vídeos, además en la parte derecha se compara el sistema

[1]http://www. parcan.es/video/canales.py

propuesto haciendo uso de tres técnicas. El método básico, referido como Base, que solo hace uso del detector de caras para localizar al interviniente. En segundo lugar, la diarización basada en clasificación de planos (DSC) que combina el clasificador de planos y la detección de personas que están de pie en los planos medios. Y en tercer lugar, el sistema completo, diarización basada en clasificación de planos y detección de fotogramas claves (DSCK) que usa el método DSC y la verificación biométrica de la cara.

Para entrenar el clasificador de planos se han usado cuatro vídeos cuyos fotogramas han sido etiquetados manualmente en cuatro clases: plano largo (4,740 muestras), plano medio (1,395 muestras), primer plano (4,309 muestras) y otros planos con 143 muestras que representan títulos y votaciones. La asignación de los umbrales para los coeficientes de la Ecuación 4.2 ha sido calculada usando los vídeos del Parlamento. Se empleó la mediana sobre los valores de coeficientes calculados y se aplicó un intervalo de ± 3.

En la evaluación del sistema se emplean las medidas descritas en [Cong et al., 2010], las medidas True Re-identification Rate (TRR) y True Distinction Rate (TDR). TRR evalúa como de bueno es el método re-identificando a los individuos, mientras TDR evalúa como de bueno es el método distinguiendo entre los individuos. Ambas medidas son formuladas como se muestra a continuación:

$$TRR = \frac{\mathrm{tr}(score)}{N} \tag{4.3}$$

$$TDR = 1 - \frac{(score\ \mathbf{1}_N)^T\ \mathbf{1}_N - \mathrm{tr}(score)}{N(N-1)} \tag{4.4}$$

donde $\mathbf{1}_N$ es el vector de dimensión N con todos los elementos con valor uno; y $\mathrm{tr}(score)$ es la traza de $score$. Siendo $score$ la matriz de $N \times N$ que tiene el resultado de la comparación de la identidad de cada individuo propuesto respecto a todas las identidades de los individuos, 1 es asignado a identidades iguales y 0 a las diferentes. Entonces, 1 en los elementos diagonales y 0 en elementos fuera de la diagonal componen una puntuación perfecta.

Los resultados se muestran en la Tabla 4.1. En general, las técnicas DSC y DSCK proporcionan mejores resultados que el método Base. Este hecho puede ser debido a que el método Base tiene muchos fallos a la hora de detectar a los usuarios en los planos medios porque detecta como intervinientes a las personas que le rodean. Un ejemplo claro se observa en el vídeo 2959, el mayor número de planos son del tipo de planos medio, mejorando los resultados en lo que refiere a la medida TRR en 71.4 %. Por otro lado, el uso de medidas biométricas mejora los resultados en un 100 % en tres vídeos. De forma general sobre los vídeos se obtiene un incremento medio de 4.7 % para la medida TRR usando el método DSC, mientras se mantiene el resultado

Características de los vídeos				Resultados					
ID	Imágenes	Planos	usuarios	Base		DSC		DSCK	
2770	314050	660	8	100.0	100.0	100.0	100.0	93.8	100.0
2785	242850	325	32	40.0	99.6	50.0	99.4	100.0	100.0
2786	265500	396	17	75.0	100.0	76.9	100.0	80.0	100.0
2787	464000	738	24	80.8	99.2	88.5	99.2	79.4	97.8
2789	232350	334	26	92.3	99.8	100.0	100.0	50.0	100.0
2790	243450	451	13	94.1	96.9	93.3	96.7	100.0	100.0
2791	442625	636	25	82.8	99.6	80.7	98.8	83.8	97.3
2792	162000	318	11	100.0	100.0	100.0	100.0	83.3	96.2
2799	241925	269	33	0.0	99.9	0.0	100.0	100.0	100.0
2800	273300	255	19	66.7	98.8	70.6	99.3	57.1	97.2
2817	299450	281	18	73.9	98.7	72.0	98.7	85.7	97.9
2818	540350	713	14	73.3	100.0	92.3	99.4	47.1	98.9
2904	247725	389	30	100.0	100.0	100.0	100.0	100.0	100.0
2905	293400	325	20	73.9	97.8	66.7	97.8	62.5	95.1
2907	210500	257	15	87.5	97.4	87.5	97.4	100.0	100.0
2908	350025	503	24	90.5	99.1	95.0	99.1	89.3	97.6
2918	122075	143	7	90.0	94.7	90.0	94.7	83.3	90.5
2940	297250	402	17	71.4	97.9	71.4	97.9	66.7	96.4
2959	217925	265	24	28.6	99.4	100.0	100.0	60.0	98.4
2960	317850	340	22	66.7	99.3	60.0	99.1	66.7	97.8
2977	247575	447	32	0.0	99.8	0.0	100.0	100.0	100.0
2978	323175	371	20	80.0	99.8	80.0	99.8	69.0	95.5
2992	192900	149	2	100.0	100.0	100.0	100.0	100.0	100.0
2995	265475	580	9	76.5	99.0	76.5	99.0	52.2	98.1
3011	182550	315	25	0.0	98.7	0.0	98.3	100.0	100.0
3012	325750	365	24	71.4	98.7	63.2	98.1	75.0	98.4
3013	382900	501	19	66.7	97.8	72.2	98.5	84.2	98.5
3014	251050	270	20	33.3	98.5	50.0	99.1	62.5	96.7
3015	274100	252	13	80.0	97.3	83.3	97.3	80.0	99.3
3017	278400	291	18	100.0	100.0	100.0	100.0	92.9	98.2
3020	332950	390	14	72.2	98.0	91.7	99.5	66.7	97.6
Media	277672	376	19	69.9	98.9	74.6	**98.9**	**79.7**	98.2
Mediana	273300	340	19	75.0	99.2	80.7	**99.2**	**83.3**	98.4

Tabla 4.1: Características principales de los vídeos y resultados para estos empleando diferentes técnicas.

para TDR. Comparando las técnicas DSC y DSCK, se obtiene una mejora media de 5.1 % en TRR. Finalmente, el incremento en TRR del método DSCK con respecto a la técnica base es de un 9.8 %.

Figura 4.6: Los fragmentos de audio pueden incluir diferentes planos visuales de personas.

Asimismo, se comprobó que las anotaciones proporcionadas por el Parlamento no eran adecuadas para comprobar el rendimiento en la re-identificación de intervinientes usando solo características visuales. Esto es debido a que las anotaciones estaban basadas en la duración de la intervención del interviniente, lo que no asegura que las cámaras estén captando imágenes de esta persona. El equipo de producción decide a que personas poner el foco de atención durante la intervención del interviniente, lo cual suscita un gran problema a la hora de usar técnicas basadas en visión por computador porque las cámaras pueden estar grabando a otra persona que no es el interviniente.

Así que para una anotación de audio específica, fragmento de audio, pueden aparecer distintos planos de diputados diferentes en la secuencia de vídeo, como se muestra en la Figura 4.6. Por esta razón, proponemos el uso de cuatro técnicas de selección de planos representativos:

- Primera aparición (FA): la persona del primer plano detectado por el sistema en el fragmento de audio es escogida como el interviniente representativo del plano.
- El más frecuente (MF): la persona que el sistema detecta con mayor número de ocurrencias en el fragmento de audio es tomada como interviniente representativo del plano.
- El más largo (GL): la persona que el sistema detecta con el plano de mayor duración en el fragmento de audio es escogida como interviniente representativo del plano.

- El más largo total (GTL): la persona que el sistema detecta con la mayor duración en el fragmento de audio es tomada como interviniente representativo del plano.

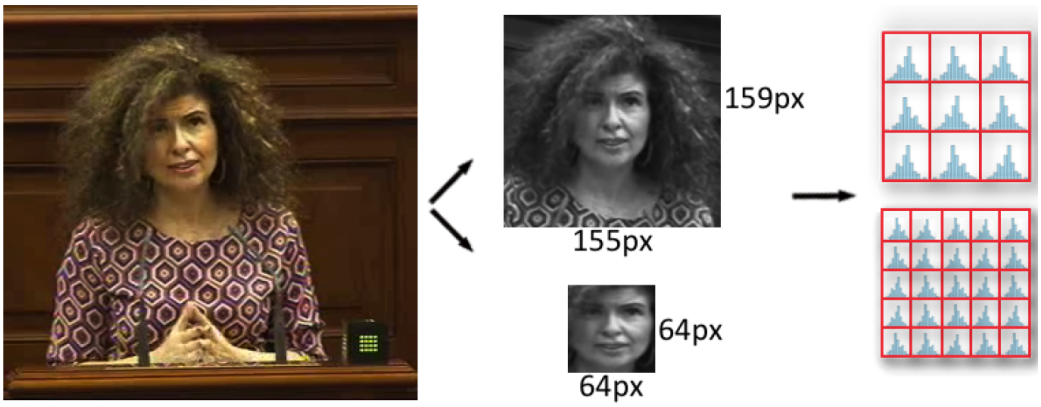


Figura 4.7: La imagen es normalizada usando el patrón cara o HS. Luego, esta es dividida en una rejilla de 3×3 o 5×5 donde un descriptor local es aplicado.

Además, a modo de complementar los resultados mostrados anteriormente, se han empleado dos regiones de interés con el fin de determinar las medidas de distancia y descriptores locales que rinden mejor en este escenario. El patrón Cara y el patrón Cabeza-Hombro (HS), este último abarca la zona de la cabeza y los hombros. En su interior se obtienen los descriptores locales usando dos tipos de rejillas $(3 \times 3, 5 \times 5)$ estas configuraciones se han escogido en base a los buenos resultados suscitados en [Castrillón-Santana et al., 2017] aplicando estas áreas de interés para la aplicación de descriptores locales. Un ejemplo del proceso de modelado se muestra en la Figura 4.7.

Como descriptores locales se han empleado los siguientes: Histogramas de Gradientes Orientados (HOG) [Dalal and Triggs, 2005], Patrones Binarios Locales (LBP) [Ojala et al., 1994], LBP Uniforme (LBPu2) [Ojala et al., 2002], LBP basado en Intensidad (NILBP) [Liu et al., 2012], Patrones de Gradientes Locales (LGP) [Jun and Kim, 2012], Cuantificación de Fase Local (LPQ) [Ojansivu and Heikkilä, 2008], Patrones Locales Salientes (LSP0) [Chai et al., 2013], Patrones Ternarios Locales (LTP) [Tan and Triggs, 2010], Patrones Ternarios Locales altos (LTPh) [Tan and Triggs, 2010], Patrones Ternarios Locales bajos (LTPl) [Tan and Triggs, 2010], Descriptor Local de Weber (WLD) [Chen et al., 2010] y Amplificador de Información de Estadísticas Orientadas Localmente (LOSIB) [García-Olalla et al., 2014]. Además, se usan las distancias de Canberra, Chebyshev, Coseno, Euclídea y divergencia KL para la medida de histogramas.

Para hacerse una idea del coste de la ejecución de los experimentos, para cada vídeo se han ejecutado dos tipos de regiones de interés con dos tipos de rejillas con 12 descriptores locales aplicando cinco medidas de distancia diferentes. Asimismo, se han usado cuatro enfoques para la selección de planos representativos, alcanzando 960 experimentos por vídeo.

Dada la gran cantidad de resultados que se han obtenido, la medida F es empleada para obtener solo una medida que proporcione una compensación entre TRR y TDR. En este caso se define con la siguiente ecuación:

$$F = 2 \cdot \frac{TRR \times TDR}{TRR + TDR}$$

con lo que se consigue que si obtenemos un 0 % en TRR y 100 % en TDR, se obtenga un resultado del 0 %.

Centrándonos en las técnicas para asignar los planos representativos al fragmento de audio, se ha calculado la medida F media de todos los vídeos procesados. También, el valor medio de los diferentes descriptores locales y medidas de distancia, ver Tabla 4.2. La asignación del interviniente más frecuente alcanza los valores más altos independientemente del tipo de región de interés y de la rejilla empleada.

Técnica	Cara		HS	
	3×3	5×5	3×3	5×5
FA	56.61	52.23	64.03	59.51
MF	56.70	54.91	**64.31**	59.57
GL	56.19	54.22	63.85	59.05
GTL	54.21	54.65	61.70	57.65

Tabla 4.2: Comparación de diferentes patrones y número de celdas respecto a técnicas de selección de planos representativos en términos de medida F para el valor medio de todos los vídeos procesados, descriptores y distancias.

En la Tabla 4.3 se muestra a la comparación de diferentes descriptores locales con las regiones de interés y las rejillas. El mejor descriptor es WLD obteniendo un incremento de 0.39 % en relación con el segundo mejor descriptor, HOG. El primero obtiene una mejora de 5.86 % en relación con el peor descriptor de la configuración.

En relación con las medidas de distancia de histogramas, Canberra es la distancia que alcanza el valor máximo como se puede observar en la Tabla 4.4. Pero en general, la divergencia KL tiene un buen rendimiento en las diferentes configuraciones y la diferencia entre la distancia de Canberra y esta es pequeña.

Descriptor	Cara		HS	
	3×3	5×5	3×3	5×5
HOG	56.81	55.09	65.86	63.51
LBP	55.03	53.17	62.10	58.36
LBPu2	55.95	55.52	63.93	58.97
LGP	53.52	51.72	64.58	59.41
LOSIB	49.56	47.57	64.72	60.65
LPQ	58.75	53.19	60.62	55.65
LSP0	55.49	54.70	59.25	53.75
LTPh	56.87	54.35	62.42	58.79
LTPl	56.40	54.05	63.29	57.02
LTP	56.97	54.65	62.75	59.77
NILBP	56.60	56.66	65.91	57.63
WLD	59.20	57.34	**66.25**	63.83

Tabla 4.3: Comparación de diferentes patrones y número de celdas respecto a descriptores locales en términos de la medida F para el valor medio de los vídeos procesados, técnicas de selección de interviniente representativo y distancias.

Distancia	Cara		HS	
	3×3	5×5	3×3	5×5
Canberra	54.13	53.53	**65.86**	62.16
Chebyshev	52.84	47.25	55.76	46.81
Coseno	57.58	55.94	65.04	62.50
Euclídea	58.74	55.86	65.16	60.07
KL	56.35	57.43	65.54	63.18

Tabla 4.4: Comparación de diferentes patrones y número de celdas respecto a medidas de distancia en términos de la medida F para el valor medio de todos los vídeos, técnicas de selección de interviniente representativo y distancias.

Adicionalmente, destacamos que el uso del patrón HS y la rejilla de 3×3 generan mejores resultados en todos los experimentos. En particular, la mejor configuración presenta un 74.09 % usando la técnica de selección de interviniente MF, empleando el patrón HS con una rejilla de 3×3 con el descriptor WLD y realizando el emparejamiento usando la distancia de Canberra.

4.2. Detección de la novedad en re-identificación

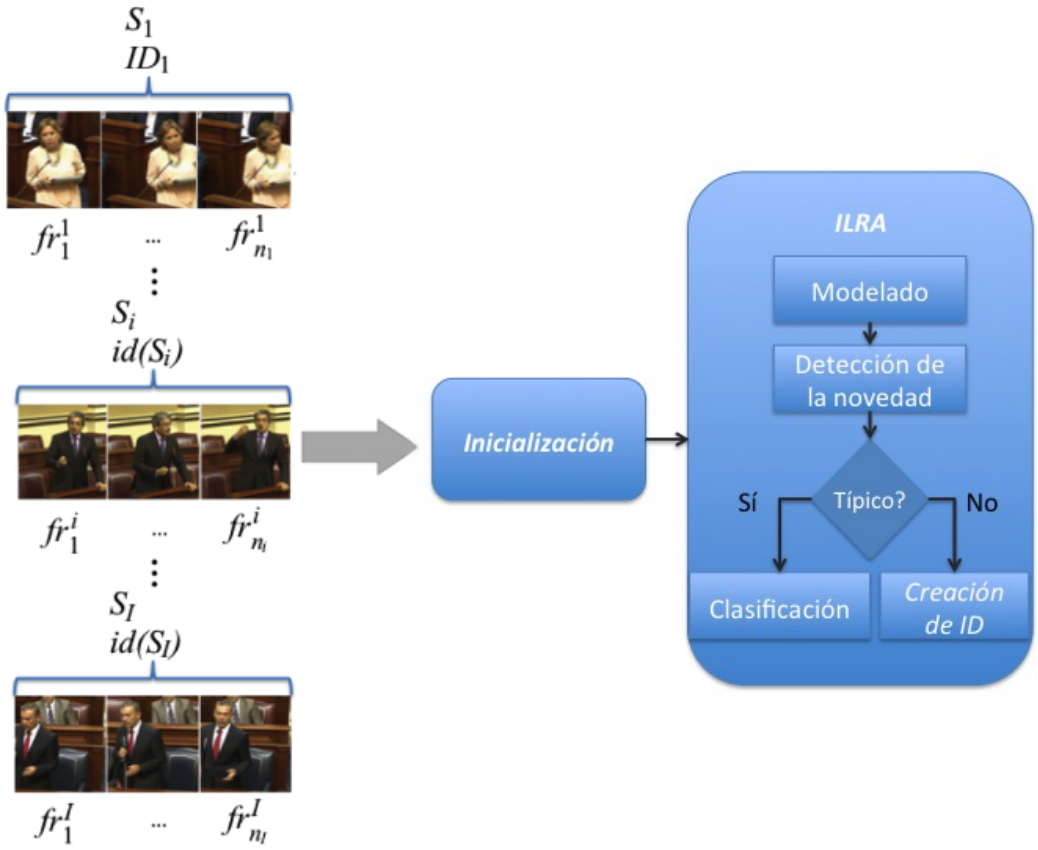


Figura 4.8: Un vídeo es dividido en planos, S_i, que están compuestos por imágenes, fr_i. Estos alimentan al sistema propuesto que está formado por dos etapas. La fase de inicialización y la fase ILRA.

4.2.1. Escenario

Cuando la re-identificación de personas tiene lugar en un escenario de mundo abierto, como en el caso del parlamento descrito anteriormente, hay que tener en cuenta la situación de añadir una persona no registrada como una nueva identidad al sistema. Esta inclusión puede producir errores que van a afectar al rendimiento del sistema hasta que finalice el proceso de re-identificación, por lo que es importante hacer hincapié en esta fase del

proceso de reconocimiento de personas. En esta sección vamos a entrar en detalle en una técnica desarrollada para mejorar la detección de nuevas identidades y que no está basada en la elección de un umbral como se describió en la Sección 4.1.2.5. Se parte del escenario planteado en la sección anterior, las sesiones parlamentarias. El sistema está formado por dos etapas. La fase de inicialización se lleva a cabo sin el proceso de modelado, a diferencia de la segunda fase, ILRA (acrónimo en inglés de Isometric Log-Ratio transformation of A posteriori probabilities).

Previamente a la fase de inicialización, el vídeo es preprocesado manteniendo solo las imágenes que contienen caras frontales. Esto se lleva a cabo siguiendo el mismo proceso desarrollado en la Sección 4.1. De esta forma, el vídeo está formado por una secuencia de I planos $(S_1, \ldots, S_I)$ donde un plano (S_i) es definido como una secuencia de fotogramas (fr_i) donde aparece un único interviniente, ver la Figura 4.8. En la fase de inicialización, el sistema asigna la identidad ID_1 al primer plano $(K = 1)$. Los siguientes planos son procesados con el detector de novedad uno a uno, teniendo en cuenta de que en cada plano hay un solo interviniente. Además, el sistema tiene que reconocer si el interviniente del plano actual ha sido identificado previamente (típico) o no (atípico). Esta fase finaliza cuando se detecta un plano atípico, de tal modo que localizamos un interviniente atípico. Así que el sistema ha identificado a dos intervinientes $(K = 2)$.

Una vez el sistema ha registrado dos intervinientes, los siguientes planos son procesados para resolver nuevamente un problema de detección de novedad. Se emplea un modelado novedoso basado en las probabilidades a posteriori de los individuos procesados previamente. Este modelado no puede ser llevado a cabo en la fase anterior porque el sistema necesita tener registrado al menos dos intervinientes. Si el plano nuevo es típico, un clasificador de K-etiquetas se usa para reconocer cual de los intervinientes conocidos corresponde al del plano actual, en caso contrario una nueva identidad se asigna al plano actual. Este proceso se repite hasta que no queden más planos en la secuencia $S_1, \ldots, S_I$. En las siguientes subsecciones se describen las diferentes etapas.

4.2.2. Desarrollo

4.2.2.1. Preprocesado del vídeo

Una secuencia de vídeo se compone de $S_1, \ldots, S_I$ planos y cada plano S_i está compuesto por $fr_1^i, \ldots, fr_{n_i}^i$ imágenes con una cara detectada, en el caso de que aparezcan múltiples rostros, se selecciona la cara más grande. Para cada plano S_i del vídeo se obtiene una matriz $X_i = [\mathbf{x}_1^i, \ldots, \mathbf{x}_{n_i}^i]^\intercal$, siendo

n_i el número de fotogramas del plano i-ésimo. Las caras detectadas de cada fotograma, fr_j^i, se representan con descriptor calculado sobre la región de la rostro. Además, cada fila $\mathbf{x}_j^i$ de la matriz X_i corresponde al descriptor de dimensión D, $\mathbf{x}_j^i = desc(fr_j^i) \in \mathbb{R}^D$ $(j = 1, \ldots, n_i, \quad i = 1, \ldots, I)$, resultando una matriz de dimensión $n_i \times D$:

$$X_i = \begin{pmatrix} x_{11}^i & \cdots & x_{1D}^i \\ \vdots & \ddots & \vdots \\ x_{n_i1}^i & \cdots & x_{n_iD}^i \end{pmatrix} \tag{4.5}$$

4.2.2.2. Inicialización

En primer lugar, el sistema asigna la identidad ID_1 al primer plano S_1, obteniendo una matriz extendida que incluye la etiqueta del interviniente del plano:

$$X_1^e = \left(\begin{array}{ccc|c} x_{11}^1 & \cdots & x_{1D}^1 & ID_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n_i1}^1 & \cdots & x_{n_iD}^1 & ID_1 \end{array} \right) \tag{4.6}$$

A partir de ahora nos referiremos como identidades a las etiquetas (ID_x) dadas por el sistema al registrar a un individuo. Luego, el sistema tiene que determinar la identidad del interviniente en los siguientes planos hasta que encuentra el primer plano atípico. Esta fase tiene similitudes con la estrategia OVO (acrónimo en inglés de One Vs One) [Knerr et al., 1990] porque hasta este momento el sistema solo conoce un interviniente. Por lo tanto, el procedimiento tiene que detectar si el interviniente en el próximo plano es el mismo individuo ID_1 (típico) o si él/ella es distinto (atípico). En términos de un problema de clasificación, una máquina de vectores de soporte (SVM) [Cortes and Vapnik, 1995] es entrenada con las matrices extendidas $X_1^e, \ldots, X_{i-1}^e$, obteniendo las predicciones para la matriz de entrada X_i.

En este sentido, para cada imagen en X_i se va a generar una predicción en términos de típico/atípico. Sin embargo, no todos los fotogramas necesariamente han sido predichos con la misma etiqueta; siendo razonable considerar el plano completo S_i como típico ($id(S_i) = ID_1$) si la mayoría de los n_i fotogramas en el plano S_i son predichos como típicos, si no, como atípicos ($id(S_i) = ID_2$), incrementando el número de intervinientes K. Por lo tanto, hemos decidido utilizar el principio WTA (acrónimo en inglés de Winner-Takes-All) para este propósito.

4.2.2.3. ILRA

Una vez que el sistema ha registrado al menos dos individuos ($K \geq 2$), es necesario determinar si el individuo del siguiente plano S_i está registrado o no. Para este propósito, esta etapa comprende tres procesos principales: modelado, detección de novedad y, si el plano actual es típico, clasificación. Esta etapa tiene similitudes con una estrategia OVA (acrónimo en inglés de One Vs All) [Clark and Boswell, 1991]. Los datos disponibles en esta etapa son, por un lado, las matrices extendidas $X_1^e, \ldots, X_{i-1}^e$ que son los descriptores de los fotogramas de cada uno de los planos anteriores añadiéndole la etiqueta de sus respectivas identidades asociadas, y por otro lado, los descriptores de los fotogramas del plano actual S_i, es decir, X_i.

El objetivo de la etapa de modelado es obtener la probabilidad *a posteriori*, $p_{jk}^i = \text{Prob}(ID_k|\mathbf{x}_j^i)$, de cada fotograma j en el plano S_i de pertenecer a cada identidad registrada k. Además, para el plano i-ésimo se calcula la matriz P_i:

$$P_i = \begin{pmatrix} p_{11}^i & \cdots & p_{1K}^i \\ \vdots & \ddots & \vdots \\ p_{n_i1}^i & \cdots & p_{n_iK}^i \end{pmatrix} \tag{4.7}$$

donde $\sum_{k=1}^{K} p_{jk}^i = 1$. Por un lado, los planos $S_1, \ldots, S_{i-1}$ donde la identidad ha sido asignada, la estimación de las probabilidades *a posteriori* se realiza empleando la estrategia *leave-one-out*. Por lo tanto, para cada fotograma $fr_j \in \{S_1, \ldots, S_{i-1}\}$, las probabilidades *a posteriori* se calculan usando un clasificador bayesiano entrenado con todas las imágenes menos el fotograma fr_j, $\{S_1, \ldots, S_{i-1}\} \setminus fr_j$. Por la otra parte, para cada fotograma $fr_j \in S_i$, las probabilidades *a posteriori* se calculan usando el clasificador bayesiano entrenado con todas las imágenes de los planos anteriores, $\{S_1, \ldots, S_{i-1}\}$.

Una vez que las probabilidades *a posteriori* han sido calculadas, se lleva a cabo el segundo paso del proceso de modelado. Se aplica la transformación ILR (acrónimo en inglés de Isometric Log-Ratio) a las matrices $P_1, \ldots, P_i$. Esta es una transformación bien conocida en el campo de datos de composición que obtiene una representación de coordenadas real, preservando la métrica de Aitchison en el espacio original de las probabilidades *a posteriori* [Egozcue et al., 2003]. Formalmente es definida como:

$$Z_i = ilr_v = clr(P_i)V \tag{4.8}$$

donde *clr* es la transformación CLR (acrónimo en inglés de Centered Log Ratio) y V es una matriz cuyas columnas forman una base ortonormal del plano CLR. Como resumen, cada fotograma j-ésimo se normaliza de la siguiente

manera:

$$\mathbf{x}_j^i \in \mathbb{R}^D \quad \Rightarrow \quad \mathbf{p}_j^i \in \mathbb{R}^K \quad \Rightarrow \quad \mathbf{z}_j^i \in \mathbb{R}^{K-1} \tag{4.9}$$

Entones, todos los vectores transformados están organizados en filas en la matriz Z_i y esta es la matriz que caracteriza el plano S_i para determinar la identidad del interviniente. Se sigue un procedimiento de transformación similar para todos los fotogramas en los planos $S_1, \ldots, S_{i-1}$, obteniendo las matrices $Z_1, \ldots, Z_{i-1}$. Para determinar la novedad en el plano S_i, un clasificador SVM de una clase es entrenado con las matrices extendidas $Z_1^e, \ldots, Z_{i-1}^e$, y de forma similar a la detección de la novedad de la fase de inicialización, se obtienen las predicciones para la matriz de entrada Z_i. Nuevamente, se usa la estrategia WTA para determinar si S_i es atípico o típico. En el primer caso, $id(S_i) = ID_{K+1}$ es asignado y el número de identidades conocidas por el sistema es incrementado. En el otro caso, cuando S_i es considerado típico, se utiliza un clasificador para identificar a cuál de los conocidos pertenece. El módulo de clasificación puede ser desarrollado por cualquier tipo de clasificador, el cual es entrenado con las matrices extendidas $Z_1^e, \ldots, Z_{i-1}^e$ para determinar $id(S_i)$. Además, la estrategia WTA es escogida para determinar la identidad que caracteriza al plano S_i.

4.2.3. Experimentos y resultados

Identificador del vídeo	Intervinientes	Planos	fotogramas	Duración
2771	5	13	2,440	0:33:23
2918	7	33	7,142	1:21:23
3015	8	52	22,088	3:02:44
2792	11	55	13,956	1:48:00
2907	12	57	9,542	2:20:20
3011	21	73	6,525	2:01:42

Tabla 4.5: Descripción de los vídeos analizados. Las columnas, "planos" y "fotogramas" indican el número de planos y fotogramas.

Con el fin de evaluar esta propuesta, se han escogido nuevamente las grabaciones del Parlamento de Canarias. Para los experimentos se escogieron seis vídeos con diferentes características, las cuales aparecen sintetizadas en la Tabla 4.5. Los vídeos seleccionados cubren un amplio rango de intervinientes (5-21) y planos, por tanto, se puede evaluar la influencia del número de intervinientes. Los planos con una duración inferior a 30 segundos son obviados ya que se consideran planos irrelevantes para las aplicaciones de diarización. Además, los fotogramas donde no aparecen caras son descartados. Para este fin se emplea un detector de rostros basado en un clasificador

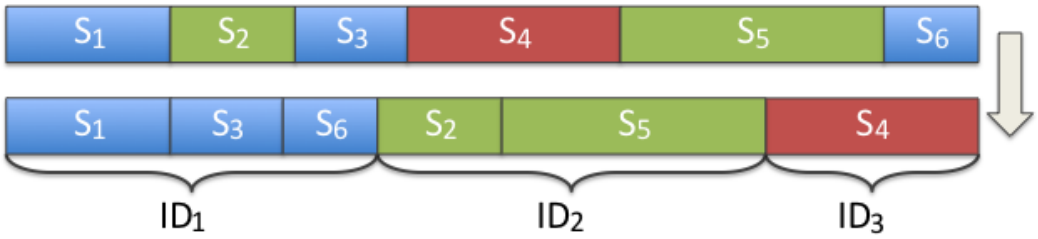


Figura 4.9: Los planos originales son reorganizados con el propósito de formar grupos por ID para los experimentos de detección de novedad (inicialización y ILRA) y clasificación (ILRA).

SVM con características HOG [Kazemi and Sullivan, 2014]. Las caras son normalizadas de forma similar a la Sección 4.1 donde se tiene en cuenta el tipo de plano y los fotogramas claves. Atendiendo al número de intervinientes, los vídeos pueden ser clasificados como cortos con menos de nueve intervinientes (vídeos con los identificadores 2771, 2918, 3015) y largos con más de diez (vídeos con los identificadores 2792, 2907, 3011).

En primer lugar, un conjunto de experimentos offline han sido llevados a cabo y evaluados en diferentes situaciones que involucran a la metodología propuesta. La propuesta de evaluación comprende tres experimentos principalmente: 1) detección de la novedad en la fase de inicialización, 2) detección de la novedad y 3) clasificación en la fase ILRA. En este sentido, el rendimiento en las diferentes fases de nuestra propuesta puede ser evaluado. Con este objetivo, los planos con el mismo ID son reorganizados para ejecutar los experimentos debidamente, como se muestra en la Figura 4.9.

Como resultado del reordenamiento de las muestras, el conjunto de entrenamiento está desbalanceado porque hay IDs más presentes que otras; para evitar esta situación, 500 fotogramas son escogidos aleatoriamente por identidad. Cuando el número de imágenes por identidad es menor que quinientos, todos los fotogramas del plano se usan. Para validar el proceso se usaron 100 repeticiones.

La dimensionalidad del rostro de los individuos es reducida, $\mathbb{R}^{w \times h} \rightarrow \mathbb{R}^{D}$, como se comenta en la Sección 4.2.2. Esta reducción está basada en la aplicación de un descriptor al área de la cara del interviniente $(w \times h)$ donde w y h representan el ancho y el alto respectivamente. Dos tipos de descriptores han sido evaluados, descriptores locales y profundos. El primer tipo usa una rejilla de 3×3 celdas sobre la imagen alineada de 59×65 píxeles. Los siguientes descriptores locales son evaluados: HOG, LBP, LBPu2, NILBP, WLD con un tamaño de 81, 2304, 531, 531, 2304 respectivamente. El último tipo corresponde al vector de características extraído de una red profunda. En este caso, una arquitectura basada en tripletas usando la red

Vídeo		Descriptor	Detección de la novedad en la fase de inicialización			Detección de la novedad en la fase ILRA			Clasificación de intervinientes en la fase ILRA	
Id	K		Típico	Atípico	F	Típico	Atípico	F	MAP Acc.	SVM Acc.
2771	5	HOG	100.0	90.00	94.74	80.00	60.00	**68.57**	96.52	96.92
		LBP	80.00	90.00	84.71	40.00	60.00	48.00	62.15	64.07
		LBPu2	80.00	90.00	84.71	80.00	60.00	**68.57**	96.72	96.89
		NILBP	100.0	90.00	94.74	20.00	80.00	32.00	72.45	81.13
		Resnet_T	100.0	100.0	**100.0**	60.00	80.00	**68.57**	**98.51**	**98.05**
		WLD	80.00	100.0	88.89	40.00	80.00	53.33	94.17	94.17
2918	7	HOG	85.71	52.38	65.02	100.0	85.71	**92.31**	92.15	91.12
		LBP	85.71	85.71	85.71	85.71	57.14	68.57	44.34	47.57
		LBPu2	85.71	90.48	88.03	28.57	100.0	44.44	**98.63**	**98.03**
		NILBP	100.0	85.71	92.31	100.0	0.00	0.00	41.05	50.20
		Resnet_T	100.0	100.0	**100.0**	29.57	86.71	42.86	97.49	97.16
		WLD	85.71	80.95	83.27	42.85	85.71	57.14	92.89	92.99
3015	8	HOG	100.0	85.71	92.31	100.0	100.0	**100.0**	95.30	94.24
		LBP	87.50	100.0	93.33	25.00	37.50	30.00	54.92	56.47
		LBPu2	87.50	100.0	93.33	50.00	100.0	66.67	**97.93**	**98.01**
		NILBP	100.0	96.43	98.18	87.50	12.50	21.88	63.00	67.58
		Resnet_T	100.0	100.0	**100.0**	27.27	100.0	42.85	97.78	97.52
		WLD	100.0	96.43	98.18	87.50	0.00	0.00	63.57	68.68
2792	11	HOG	81.82	89.09	85.30	90.91	100.0	**95.24**	92.60	91.26
		LBP	90.91	98.18	94.41	18.18	72.72	29.09	51.84	53.42
		LBPu2	81.82	96.36	88.50	45.45	90.91	60.61	97.12	96.84
		NILBP	81.82	96.36	88.50	100.0	0.00	0.00	45.16	55.76
		Resnet_T	100.0	100.0	**100.0**	36.00	100.0	52.94	**97.94**	**97.83**
		WLD	81.82	98.18	89.26	9.09	90.91	16.53	85.13	85.31
2907	12	HOG	75.00	90.91	82.19	41.66	83.33	55.56	96.42	96.02
		LBP	75.00	96.97	84.58	66.67	75.00	70.59	64.30	64.47
		LBPu2	66.67	98.48	79.51	25.00	83.33	38.46	98.11	98.19
		NILBP	83.33	100.0	90.91	50.00	25.00	33.33	76.19	79.07
		Resnet_T	100.0	100.0	**100.0**	41.67	100.0	58.83	**98.90**	**98.98**
		WLD	58.33	100.0	73.68	75.00	91.67	**82.50**	92.25	91.87
3011	21	HOG	52.38	94.76	67.47	47.62	71.43	57.14	41.29	**96.55**
		LBP	42.86	97.14	59.48	61.90	61.90	61.90	20.18	49.65
		LBPu2	42.86	96.19	59.30	42.86	76.19	54.86	40.85	94.92
		NILBP	57.14	96.19	71.69	61.90	52.38	56.75	36.98	84.26
		Resnet_T	76.19	98.57	**85.95**	23.81	90.48	37.70	**41.49**	94.64
		WLD	28.57	99.05	44.35	85.71	80.95	**83.27**	36.70	86.09
Promedio		HOG	82.49	83.81	81.17	76.70	83.41	**78.14**	85.71	94.35
		LBP	77.00	94.67	83.70	49.58	60.71	51.36	49.62	55.94
		LBPu2	74.09	95.25	82.23	45.31	85.07	55.60	88.23	97.15
		NILBP	87.05	94.12	89.39	69.90	28.31	23.99	55.81	69.67
		Resnet_T	96.03	99.76	**97.66**	36.22	92.70	50.62	**88.69**	**97.36**
		WLD	63.51	95.77	79.60	56.69	71.54	48.80	77.45	86.52

Tabla 4.6: Resultados de los experimentos offline en términos de la medida exactitud para la detección de la novedad en la fase de inicialización, detección de la novedad en la fase ILRA, y clasificación de intervinientes en la fase ILRA. Los resultados comprimen la evaluación de diferente descriptores. En negrita se muestran los resultados más altos.

Inception Resnet (Resnet_T) [Schroff et al., 2015, Szegedy et al., 2017] es empleada. Principalmente, la red optimiza una función objetivo que transforma las muestras a un nuevo espacio de características, donde las muestras que pertenecen a la misma identidad están próximas y las muestras de diferentes están alejadas. Entonces, tres instancias de la red Resnet_T que comparten la misma matriz de pesos son usadas. El espacio embebido es representado por la última capa totalmente conectada, con una dimensión de 128. Resnet_T es usada debido a los excelentes resultados alcanzados en diversos tipos de problemas en la actualidad. Esta arquitectura fue entrenada en el conjunto de datos MS-Celeb-1M [Guo et al., 2016] que contiene un millón de identidades y que de esta forma podemos obtener un modelo generalizado para

extraer de las caras el vector de características. La red fue inicializada con los siguientes parámetros: mini-lotes de tamaño 90 durante 500 épocas; la tasa de aprendizaje inicial es de 0.1, y esta es decrementada en un factor de 10 después de cada 100 épocas. Además, se usa un margen de distancia entre parejas positivas y negativas de 0.2 (α).

Una vez que se define el diseño experimental, es necesario adoptar una métrica para evaluarlo, para ello, *Acc* es la exactitud (accuracy) que es usada para evaluar los experimentos offline. se define formalmente de la siguiente forma:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.10}$$

donde TP y FP es el número de verdaderos y falsos positivos respectivamente; TN y FN son los números de verdadero y falsos negativos respectivamente. La exactitud es usada para medir las detecciones de típicos y atípicos. En lugar de calcular el promedio de estos valores, se adopta la medida F para obtener una única medida que proporcione una compensación entre ambas exactitudes. Es definida formalmente como se muestra en la siguiente ecuación:

$$F = 2\frac{precision \times recall}{precision + recall} \tag{4.11}$$

donde:

$$precision = \frac{TP}{TP + FP} \tag{4.12}$$

y:

$$recall = \frac{TP}{TP + FN} \tag{4.13}$$

donde *precision* es la precisión que viene determinada por la fracción de muestras relevantes sobre las muestras recuperadas; Asimismo, *recall* es la sensibilidad que se corresponde con la fracción de muestras relevantes que han sido recuperadas sobre la cantidad total de muestras relevantes. A continuación presentamos y discutimos los resultados obtenidos en los experimentos.

4.2.3.1. Evaluación de la detección de la novedad en la fase de inicialización

El objetivo del primer experimento es evaluar la capacidad del sistema para detectar identidades novedosas cuando se conoce una única identidad, es decir, $K = 1$. La detección típica o atípica se realizó de la siguiente manera: para cada identidad ID_k se consideran sus muestras como conjunto de prueba, y para formar el conjunto de entrenamiento se consideran dos situaciones diferentes.

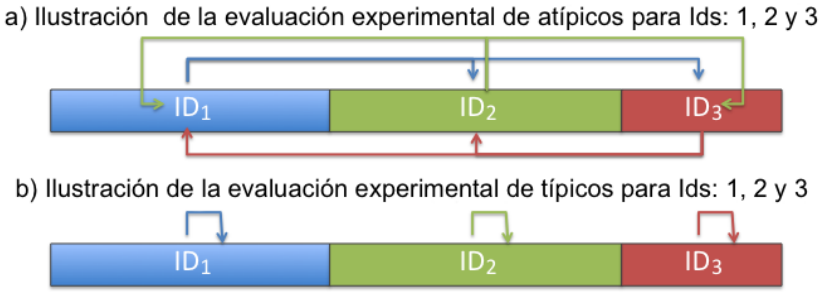


Figura 4.10: Fase de inicialización. En la imagen superior se muestra la evaluación de los experimentos de atípicos donde cada ID es emparejado individualmente con el resto de IDs (flechas coloreadas). En la imagen inferior se muestra la evaluación de los experimentos típicos donde cada ID es emparejada consigo misma (flechas coloreadas).

En el primer caso, el conjunto de entrenamiento estaba compuesto por muestras con identidades $ID_j \neq ID_k$. En tal situación, la identidad de prueba debe ser etiquetada como atípica (Figura 4.10.a) y el número de comparaciones diferentes es $K^2 - K$. Hay que tener en cuenta que, para cada comparación, la detección de los individuos tiene que ser atípica para ser un éxito. En el segundo caso, el conjunto de entrenamiento estaba compuesto por las muestras con la misma identidad ID_k. Para evitar tener la misma identidad en el conjunto de entrenamiento y de prueba, un tercio de las muestras originales de la identidad ID_k es usado como conjunto de prueba y los restantes dos tercios como conjunto de entrenamiento. En esta situación, la detección de los individuos tiene que ser típica para tener un acierto (Figura 4.10.b). Se realiza este experimento para todas las identidades ID_k en cada vídeo.

En la columna detección de la novedad de la Tabla 4.6 se resumen los resultados de los experimentos de la fase de inicialización. Se puede observar que en todos los vídeos la mejor medida F es obtenida empleando Resnet_T obteniendo un valor medio de 97.66 %. En general, los resultados de la detección de atípicos es mayor o igual al 90 % en 30 de 36 pruebas.

4.2.3.2. Evaluación de la detección de la novedad en la fase ILRA

Los experimentos offline relacionados con la fase de ILRA son motivados por la necesidad de evaluar la capacidad del método para detectar identidades novedosas de un nuevo plano cuando múltiples identidades son conocidas por el sistema. Por lo tanto, dos evaluaciones son consideradas para cada identidad, atípica y típica. La primera comprende todas las muestras de la identidad ID_k en el conjunto de prueba, mientras que el resto de identidades,

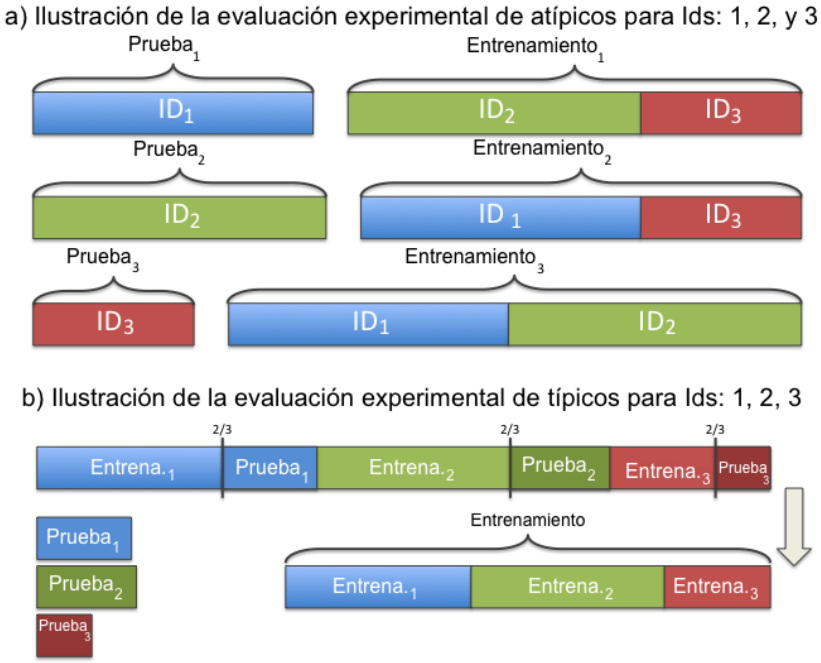


Figura 4.11: Fase ILRA. a) Evaluación experimental de atípicos donde cada ID se corresponde con los IDs restantes. b) Evaluación experimental de típicos donde cada conjunto es dividido en un tercio de prueba y el restante de entrenamiento.

$ID_{j \neq k}$, son usadas como entrenamiento (Figura 4.11.a). Este experimento se lleva a cabo para evidenciar el comportamiento del método para la detección de identidad atípica, ya que la identidad de prueba ID_k debe etiquetarse como atípica. El último comprende todas las identidades tanto en el entrenamiento como en el conjunto de prueba, dividiendo aleatoriamente y balanceando sus respectivas muestras, usando un tercio para pruebas y el resto para entrenamiento (Figura 4.11.b). Este experimento se lleva a cabo para evidenciar el comportamiento del método para la detección de identidad típica, ya que la identidad de prueba ID_k debe etiquetarse como típica.

En la columna detección de la novedad en la fase ILRA de la Tabla 4.6 alude a que el descriptor con mayor valor de F es HOG, alcanzando un 78.14 %. También se puede observar que cuando el número de intervinientes es bajo, el mejor descriptor es HOG, por el contrario, cuando el número de intervinientes es alto, WLD se comporta aparentemente mejor que el resto de descriptores.

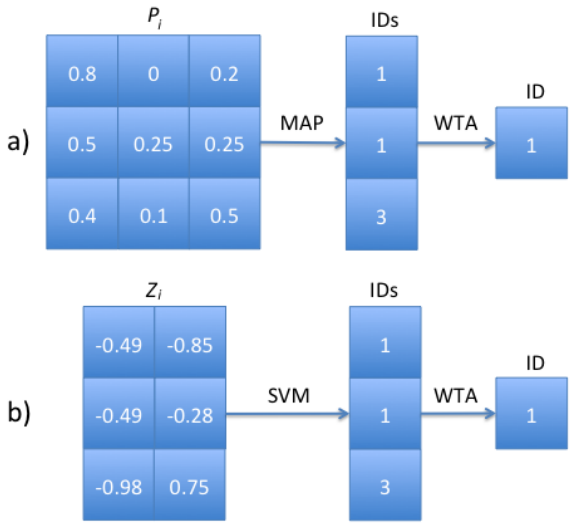


Figura 4.12: Procedimiento para determinar $id(S_i)$. a) representa el proceso para extraer la ID del interviniente usando la Máxima Probabilidad *a Posteriori* (MAP) para cada muestra. b) muestra el uso de una SVM para obtener la ID del interviniente.

4.2.3.3. Evaluación de la clasificación de intervinientes en la fase ILRA

Los experimentos offline relacionados con la etapa de clasificación se plantean para evaluar la capacidad del método para asignar correctamente la identidad de un nuevo plano donde aparece un interviniente cuando múltiples identidades son conocidas. Eso significa, cuando la identidad de un nuevo plano ($id(S_i)$) está presente sobre las identidades conocidas, el interviniente ha sido considerado como típico en la fase de ILRA, el sistema tiene que emparejarlo a la ID que le corresponde. Dos clasificadores son considerados: el clasificador por Máxima Probabilidad *a Posteriori* (MAP) extraída de las muestras (ver Figura 4.12.a) y el clasificador SVM para continuar usando la misma tipología de clasificadores que usamos a lo largo de esta propuesta (ver Figura 4.12.b). En el caso de la SVM, se utiliza un kernel RBF con $\nu = 0{,}1$, $\gamma = 0{,}1$ y $C = 1$ como principales parámetros. Se empleó una validación holdout con 100 repeticiones con remuestreo de las muestras, un tercio de estas para prueba y el resto para entrenar.

Los resultados se muestran en la columna clasificación de intervinientes de la fase ILRA de la Tabla 4.6. Entre los seis descriptores, Resnet_T alcanza la mejor exactitud en siete de los doce experimentos, proporcionando un valor medio para los clasificadores MAP y SVM de 88.69 % y 97.36 % respectivamente.

4.2.3.4. Evaluación online del sistema propuesto

ID del vídeo	Descriptor	TRR	TDR	F
2771	HOG	83.33	74.07	**78.43**
	LBP	25.00	94.44	39.53
	LBPu2	16.67	96.30	28.42
	NILBP	16.67	88.89	28.07
	Resnet_T	58.33	90.74	71.01
	WLD	8.33	96.30	15.34
2918	HOG	38.50	99.08	55.45
	LBP	95.72	11.63	20.75
	LBPu2	40.11	83.28	54.14
	NILBP	56.68	76.89	65.26
	Resnet_T	59.15	97.65	**73.68**
	WLD	31.55	93.76	47.21
3015	HOG	71.05	95.96	**81.65**
	LBP	29.47	75.76	42.44
	LBPu2	34.21	96.80	50.55
	NILBP	40.53	88.89	55.67
	Resnet_T	56.83	99.58	72.37
	WLD	36.84	96.46	53.32
2792	HOG	71.83	69.18	70.48
	LBP	28.17	85.18	42.34
	LBPu2	70.42	94.12	**80.56**
	NILBP	59.15	83.76	69.34
	Resnet_T	47.59	97.69	64.00
	WLD	54.93	82.82	66.05
2907	HOG	52.27	48.49	50.31
	LBP	15.91	87.09	26.90
	LBPu2	31.82	95.12	47.69
	NILBP	27.27	92.54	42.13
	Resnet_T	65.79	91.41	**76.51**
	WLD	40.91	74.75	52.88
3011	HOG	82.08	95.43	**88.25**
	LBP	55.66	91.34	69.17
	LBPu2	49.06	99.69	65.75
	NILBP	73.58	65.98	69.58
	Resnet_T	54.68	97.85	70.15
	WLD	71.70	89.29	79.53
Promedio	HOG	66.51	80.37	70.76
	LBP	41.66	74.24	40.19
	LBPu2	40.38	94.22	54.52
	NILBP	45.65	82.83	55.01
	Resnet_T	57.06	95.82	**71.29**
	WLD	40.71	88.90	52.39

Tabla 4.7: Resultados de los experimentos online en términos de TRR, TDR y F. En negrita se muestra la F con mayor valor.

Después de evaluar las diferentes etapas offline, realizamos los experimentos online donde se engloba el procesamiento online real. Se evalúa el mismo descriptor para cada etapa del algoritmo. El número de fotogramas por plano se ha modificado en comparación con la configuración offline. Se utilizaron 200 imágenes por plano porque el experimento comprende una mayor cantidad de planos, algunas de las cuales contienen un número reducido de fotogramas. Dado el mejor rendimiento suscitado por los clasificadores

SVM en los experimentos anteriores, SVM es adoptado para identificar los intervinientes en el caso de individuos típicos. Además, se toman TRR y TDR como medidas para evaluar los experimentos online. Asimismo, se comprimen en una única medida empleando la medida F.

Los resultados de los experimentos se resumen en la Tabla 4.7. En la mayoría de los vídeos procesados, un descriptor supera a los demás, pero no hay un comportamiento común en todos los vídeos. En este caso, el uso del descriptor depende del vídeo, no del número de intervinientes. Por un lado, destacamos la medida F obtenida en el vídeo 3011, 88.25 %, que cubre una población de 21 intervinientes y dos horas de grabación en un problema de mundo abierto, que supone un problema real y complejo. Por otro lado, el resultado obtenido con el vídeo 2907 es interesante porque presenta una deficiencia en los vectores de características tradicionales, suscitada por un problema de oclusión debido a que la mayoría de los intervinientes se colocan o se quitan las gafas durante la intervención. En esta situación, Resnet$_T$ mejora al menos un 44.69 % al resto de descriptores, alcanzando el 76.51 % en el vídeo 2907.

Además, nuestro sistema se compara con un trabajo anterior realizado por este grupo de investigación [Sánchez-Nielsen et al., 2017], por lo que sabemos, la única metodología existente en este escenario, es decir, re-identificación de intervinientes basada en la cara en sesiones de debates parlamentarios de mundo abierto. Adicionalmente, las técnicas de reconocimiento facial empleadas en mundo cerrado se utilizan para extender la comparativa con el método propuesto, ILRA. En particular, HOG, LBP LBPu2, NILBP, WLD y Resnet$_T$ son usados como vectores de características. Con el fin de detectar las muestras atípicas, se usa un umbral con valor de 0.5, siendo una muestra atípica la correspondiente a un valor mayor que el umbral. En el caso de que la muestra sea típica, un vector de distancias es calculado a partir de las muestras previamente analizadas con respecto a la muestra actual. La identidad con distancia mínima representará a la muestra actual.

ID del vídeo	Descriptor	TRR	TDR	F
2771	[Dalal and Triggs, 2005]	58.33	61.11	59.69
	[Ojala et al., 1994]	41.67	70.37	52.34
	[Ojala et al., 2002]	41.67	70.37	52.34
	[Liu et al., 2012]	33.33	79.63	46.99
	[Szegedy et al., 2017]	79.33	64.52	**71.16**
	[Chen et al., 2010]	41.67	70.37	52.34
	[Sánchez-Nielsen et al., 2017]	53.91	75.36	62.86
	Propuesto (Resnet$_T$)	58.33	90.74	71.01
2918	[Dalal and Triggs, 2005]	49.41	79.85	61.05
	[Ojala et al., 1994]	42.35	95.82	58.74
	[Ojala et al., 2002]	48.82	97.18	64.99
	[Liu et al., 2012]	57.65	85.07	68.72
	[Szegedy et al., 2017]	96.00	44.71	61.01
	[Chen et al., 2010]	43.53	94.78	59.66
	[Sánchez-Nielsen et al., 2017]	50.59	75.16	60.47
	Propuesto (Resnet$_T$)	59.15	97.65	**73.68**
3015	[Dalal and Triggs, 2005]	85.81	12.56	21.91
	[Ojala et al., 1994]	43.02	58.85	49.71
	[Ojala et al., 2002]	45.49	57.84	50.93
	[Liu et al., 2012]	48.18	51.49	49.78
	[Szegedy et al., 2017]	80.17	47.98	60.03
	[Chen et al., 2010]	69.52	38.58	49.62
	[Sánchez-Nielsen et al., 2017]	85.93	9.96	17.85
	Propuesto (Resnet$_T$)	56.83	99.58	**72.37**
2792	[Dalal and Triggs, 2005]	20.33	96.28	33.57
	[Ojala et al., 1994]	31.17	94.87	46.92
	[Ojala et al., 2002]	31.05	95.64	46.88
	[Liu et al., 2012]	48.18	51.49	49.78
	[Szegedy et al., 2017]	89.05	58.17	**70.37**
	[Chen et al., 2010]	31.17	91.56	57.27
	[Sánchez-Nielsen et al., 2017]	23.85	93.58	38.01
	Propuesto (Resnet$_T$)	47.59	97.69	64.00
2907	[Dalal and Triggs, 2005]	23.49	88.23	37.10
	[Ojala et al., 1994]	33.73	87.79	48.74
	[Ojala et al., 2002]	28.31	89.67	43.03
	[Liu et al., 2012]	26.20	88.58	40.44
	[Szegedy et al., 2017]	91.26	88.68	**89.95**
	[Chen et al., 2010]	34.94	82.92	49.16
	[Sánchez-Nielsen et al., 2017]	21.99	84.91	34.93
	Propuesto (Resnet$_T$)	65.79	91.41	76.51
3011	[Dalal and Triggs, 2005]	57.61	77.38	66.05
	[Ojala et al., 1994]	51.78	70.62	59.75
	[Ojala et al., 2002]	53.41	70.55	60.80
	[Liu et al., 2012]	58.38	73.59	65.11
	[Szegedy et al., 2017]	66.45	70.61	68.47
	[Chen et al., 2010]	50.76	78.68	61.71
	[Sánchez-Nielsen et al., 2017]	58.12	79.36	67.10
	Propuesto (Resnet$_T$)	54.68	97.85	**70.15**
Promedio	[Dalal and Triggs, 2005]	49.16	69.24	46.56
	[Ojala et al., 1994]	40.62	79.72	52.70
	[Ojala et al., 2002]	41.46	80.21	53.16
	[Liu et al., 2012]	45.32	71.64	53.47
	[Szegedy et al., 2017]	83.71	62.44	70.17
	[Chen et al., 2010]	45.27	76.15	54.96
	[Sánchez-Nielsen et al., 2017]	49.07	69.72	46.87
	Propuesto (Resnet$_T$)	57.06	95.82	**71.29**

Tabla 4.8: Resultados de los experimentos online comparados con otros métodos en términos de TRR, TDR y F. En negrita la mayor F.

El método propuesto obtiene en la mayoría de los experimentos la mejor medida F para los diferentes vídeos, comparado con los anteriores métodos. Los resultados están resumidos en la Tabla 4.8. Por un lado, el mayor incremento de los resultados es obtenido en el vídeo 3015 donde hay una mejora del 63.80 % con respecto al trabajo anterior. Siendo ILRA ampliamente superior a los métodos tradicionales de reconocimiento de caras. Por otro lado, la técnica reciente, Resnet$_T$, alcanza un incremento significativo en los resultados comparados con las técnicas clásicas. Sin embargo, el método propuesto, alcanzando una diferencia media del 1.12 % con respecto a Resnet$_T$ para los

vídeos analizados.

4.3. Interacción hombre-máquina en robots asistentes

Figura 4.13: Ejemplo de robot asistente guiando a un grupo de personas.

4.3.1. Escenario

Una vez analizados entornos controlados, con cámaras fijas, si bien pueden modificar su orientación, perspectiva y zoom se va a extender el estudio a técnicas de re-identificación en entornos no tan controlados como los del Parlamento, siendo estos entornos los relativos a los robots de servicios. Para ello, se ha desplegado un módulo de visión para la interacción entre robots colaborativos y usuarios. GidaBot [Parra et al., 2019] es un sistema para guiar a usuarios en edificios de varias plantas donde colaboran robots heterogéneos. Estos robots no pueden moverse entre plantas, dando servicio exclusivamente a una única planta, por lo que requiere la colaboración de varios robots para alcanzar el objetivo de guiar al usuario a su destino. El sistema hace uso de

comunicación entre robots para difundir los objetivos y las rutas durante el proceso de guiado.

Se plantea un módulo de re-identificación/verificación facial, evaluado de forma offline, y se integra en el sistema GidaBot. Como los usuarios son visitantes casuales, no se almacena información a largo plazo y, en consecuencia, los rostros son desconocidos para el sistema. Inicialmente, la re-identificación y la verificación se evalúan offline considerando diferentes detectores de rostros. Para el sistema real, varios detectores de rostros se fusionan en paralelo. Esta técnica de fusión supera a cualquier método individual y mejora en gran medida la fiabilidad del sistema real, como demuestran las pruebas realizadas con robots reales en la Facultad de Informática de San Sebastián (UPV/EHU).

El escenario en particular es un edificio de cuatro plantas que cuenta con dos escaleras laterales y un solo ascensor que permite a las personas moverse entre estas. La entrada principal está en la primera planta, donde se ubican algunas salas de conferencias y laboratorios. La oficina del decano, la secretaría, el auditorio y laboratorios se encuentran en la segunda planta. Mientras que los despachos de los profesores se encuentran en el tercer y cuarto piso. Así mismo, los laboratorios de investigación también se encuentran en la última planta.

4.3.2. Desarrollo

El proceso de reconocimiento facial consta de varios pasos (ver Figura 4.14). Una vez capturada la imagen, se aplica la detección de rostros y la estimación de los puntos de referencia faciales. Las ubicaciones de los ojos sirven para realizar la alineación del rostro, solo se procesan las caras cuyas marcas faciales son detectadas. En los experimentos que se describen a continuación, se usaron dos detectores de rostros diferentes, primero se usa el más rápido y el segundo solo si no se obtiene.

En escenarios con restricciones de hardware, se debe adoptar un detector de rostro sencillo y ligero. Un posible enfoque en tales escenarios es la combinación de detectores faciales como Viola-Jones [Viola and Jones, 2004], DLIB [King, 2009] o MTCNN (acrónimo del inglés de Multi-task Cascaded Convolutional Networks) [Zhang et al., 2016] (ordenados de menor a mayor requisitos hardware). Generalmente, para cualquier detector, las ubicaciones de los ojos y la nariz se utilizan para realizar la normalización, incluida una alineación de la cara con la extracción de esta región. Este proceso implica una transformación afín para su posterior rotación y escalado, obteniendo una imagen final recortada de 160×160 píxeles.

Una vez que la cara detectada es normalizada, se emplea como represen-

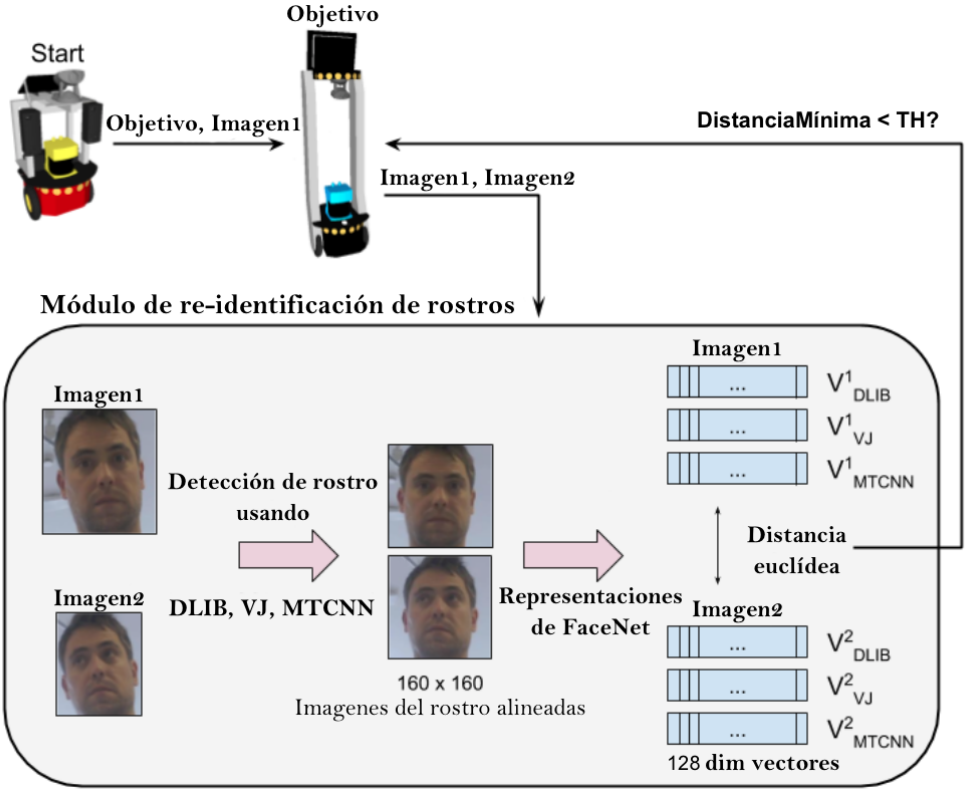


Figura 4.14: Módulo de re-identificación de usuarios en GidaBot.

tación el vector de características extraído de FaceNet [Schroff et al., 2015]. Este asigna imágenes faciales a un espacio euclídeo, donde las distancias sirven como medida de similitud facial. Este enfoque ha sido ampliamente utilizado recientemente [Amos et al., 2016, Ding et al., 2017, Golla and Sharma, 2018], encontrándose entre los más avanzados con respecto al reconocimiento facial. La ventaja que proporciona es un vector de características compacto de solo 128 valores donde se puede aplicar la distancia Euclídea para realizar comparaciones.

El proceso de entrenamiento de FaceNet se basa en la función de pérdida *triplet loss* que compara una entrada base (referencia) con respecto a una entrada positiva (identidad igual al de referencia) y la entrada negativa (identidad diferente al de referencia). Así, el conjunto de entrenamiento estará compuesto por tripletas $(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n)$, siendo $\mathbf{x}_i^a$ la muestra de referencia y $\mathbf{x}_i^p$ y $\mathbf{x}_i^n$ con la misma clase que la referencia (positiva) y diferente clase que la referencia (negativa), respectivamente. Cada muestra es la entrada de cada una de las tres ramas que configuran FaceNet y que comparten los mismos pesos. La tripleta tiene como objetivo minimizar la distancia del de referen-

cia y las entradas positivas y maximizar la distancia entre la de referencia y las entradas negativas. Formalmente, la función objetivo triplet loss se define con la siguiente expresión:

$$Loss_T = \sum_{i=1}^{N} [||f_i^a - f_i^p||_2^2 - ||f_i^a - f_i^n||_2^2 + \alpha]_+ \tag{4.14}$$

donde α es un parámetro de margen y f_i^a, f_i^p, f_i^n son los vectores de características extraídos de las tres ramas de la red, la de referencia y las entradas positivas y negativas, respectivamente, para la i-ésima tripleta. La capa superior de las ramas es una capa completamente conectada de tamaño 128 que representa el vector de características de salida.

En nuestro caso particular, FaceNet[2] ha sido entrenado con el conjunto de datos MS-Celeb-1M [Guo et al., 2016], esta ha obtenido una exactitud de 99.6 % en el conjunto de datos LFW (acrónimo en inglés de Labeled Faces in the Wild) [Shi and Jain, 2019]. De hecho, MS-Celeb-1M cubre una amplia población de personas, conteniendo un millón de rostros de 100,000 usuarios. Dado que en un escenario de robots de asistencia, estamos tratando con usuarios no incluidos en el sistema, MS-Celeb-1M enriquece el espacio facial, proporcionando un conjunto de datos generalizados que se ajusta a los requisitos de un contexto con variabilidad facial impredecible para identidades desconocidas [Tan and Triggs, 2010].

4.3.3. Experimentos y resultados

En esta sección se exponen los experimentos offline y online. Asimismo, para la ejecución de los experimentos online es necesario integrar el módulo de visión en los robots.

4.3.3.1. Rendimiento offline del módulo de re-identificación facial

Para evaluar el rendimiento del reconocimiento facial, se construyó un conjunto de imágenes con 18 voluntarios de la Facultad de Informática de la UPV/EHU. Se utilizaron cuatro robots que se ubicaron en pisos diferentes. Debido a la morfología de los robots, aunque todos los robots usaban el mismo modelo de cámara (cámara web Logitech C920 HD PRO) con una resolución de 800×600, las perspectivas de las vistas de los robots difieren entre si, por lo tanto, las imágenes variaban considerablemente. La Figura 4.15 permite comparar visualmente las diferentes perspectivas de los robots.

[2]github.com/nyoki-mtl/keras-facenet

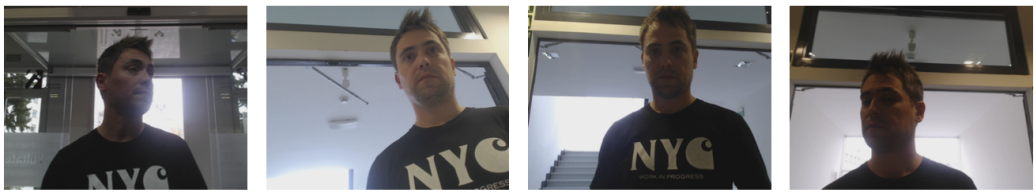


Figura 4.15: Perspectivas de las vistas de los diferentes robots de un mismo individuo.

Cada uno de los usuarios se coloca enfrente a cada robot y en diferentes ubicaciones. Se permitió a los usuarios moverse libremente, pero con el único requisito de que la cabeza permaneciera visible en la pantalla. La Figura 4.16 muestra algunas imágenes correspondientes a un solo usuario en una ubicación concreta.

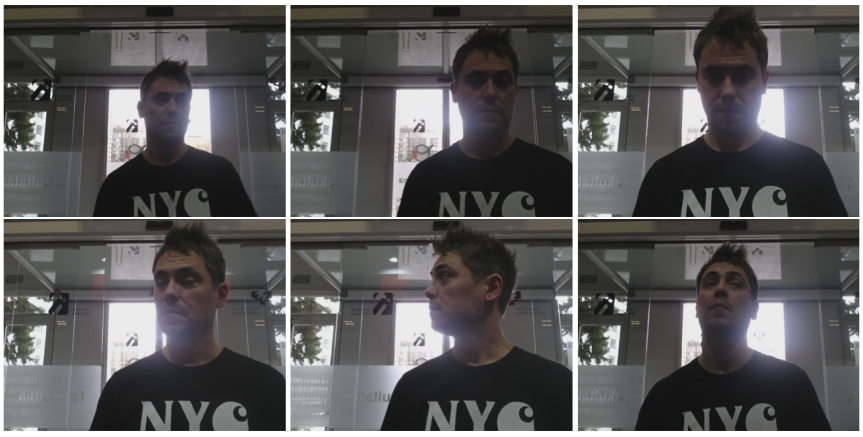


Figura 4.16: Imágenes del proceso de captura. La ubicación corresponde a la entrada principal de la facultad, donde las condiciones de iluminación son críticas.

Aunque las imágenes se capturaron a 0.5Hz, no se siguió ningún procedimiento de captura para asegurar un número idéntico de muestras por identidad. La pose de la cara no estaba restringida como se comentó anteriormente; por lo tanto, el conjunto de datos presenta una serie de desafíos para los detectores faciales. El conjunto de datos total contiene 1,808 imágenes de los usuarios, distribuidas según se muestra en la Tabla 4.9. En cada imagen, se aplicó la configuración del detector facial descrita en la sección anterior, obteniendo una detección positiva con puntos de referencia faciales en 1,316 imágenes. El número de rostros detectados en relación con las identidades y los pisos también se indica entre paréntesis en la Tabla 4.9. En la Figura 4.17 se muestran rostros normalizado de seis usuarios capturados en

	Plantas			
Id	1	2	3	4
1	87 (77)	29 (28)	22 (9)	45 (34)
2	41 (38)	16 (12)	19 (10)	35 (35)
3	27 (26)	22 (17)	26 (20)	21 (14)
4	35 (31)	21 (11)	20 (0)	27 (3)
5	19 (18)	18 (12)	24 (2)	37 (0)
6	22 (22)	6 (4)	21 (15)	41 (13)
7	25 (23)	14 (12)	21 (18)	16 (15)
8	27 (27)	20 (12)	22 (21)	36 (28)
9	24 (23)	22 (17)	20 (12)	26 (16)
10	16 (16)	24 (21)	21 (20)	26 (26)
11	32 (19)	44 (24)	22 (2)	31 (0)
12	22 (22)	23 (21)	22 (10)	21 (18)
13	26 (20)	12 (12)	21 (20)	28 (21)
14	25(25)	25 (19)	20 (15)	31 (25)
15	17 (17)	24 (22)	26 (22)	26 (23)
16	27 (27)	24 (24)	21 (15)	25 (12)
17	25 (24)	24 (19)	27 (14)	28 (19)
18	20(19)	10 (8)	13 (6)	25 (17)

Tabla 4.9: Distribución de las muestras en el conjunto de datos por identidad y planta (entre paréntesis se muestra el número final de caras detectadas por identidad y planta).

la planta 1, y además, en la Figura 4.18 se presentan tres usuarios normalizados en las cuatro plantas. Las muestras plantearon una serie de situaciones desafiantes dadas las variaciones en términos de calidad de imagen, pose e iluminación.

Figura 4.17: Ejemplos de caras detectadas que han sido normalizadas, para diferentes usuarios, después de la alineación y recorte de las muestras en la planta 1.

Rank-1	**Rank-2**	**Rank-3**	**Rank-4**	**Rank-5**	**Rank-6**	**Rank-7**	**Rank-8**	**Rank-9**	**Rank-10**
97.42	97.95	98.33	98.56	98.64	98.71	98.71	98.79	98.94	99.01

Tabla 4.10: Rank-1 a Rank-10 para los usuarios en todas las plantas del edificio.

Para evaluar el sistema, adoptamos un enfoque basado en la problemática de re-identificación. Asumimos por simplicidad un escenario de mundo cerrado donde el probe siempre estuvo presente en el gallery. De acuerdo con la literatura de re-identificación, la CMC (acrónimo en inglés de Cumulative Match Characteristics) se emplea comúnmente para evaluar las diferentes metodologías. CMC relaciona el rango con la tasa de identificación.

En un primer experimento global, se consideró todo el conjunto de datos. Cada aparición del usuario se tomó como probe y el resto del conjunto constituye el gallery. Los resultados de este experimento se muestran en la Tabla 4.10. La precisión obtenida para el Rank-1 y el Rank-10 fue respectivamente de 97.42 % y 99.01 %. Dado que el gallery estaba compuesta por

Figura 4.18: Muestras después de la normalización facial capturadas en diferentes plantas que evidencian variaciones en la pose e iluminación en las plantas. Cada columna corresponde a un participante específico (Identidades 1, 2 y 3) por planta.

18 identidades diferentes, la suposición aleatoria de rango 1 reportaría una precisión del 5.6 %. Por lo tanto, la precisión obtenida fue prometedora. Aunque este experimento proporcionó una idea sobre el rendimiento del módulo de reconocimiento facial, fue altamente optimista porque el gallery contenía imágenes del probe tomadas por el mismo robot, lo que no es la situación típica esperada en el escenario propuesto, identificar a una persona en plantas diferentes donde las condiciones lumínicas y la pose de las personas difieren.

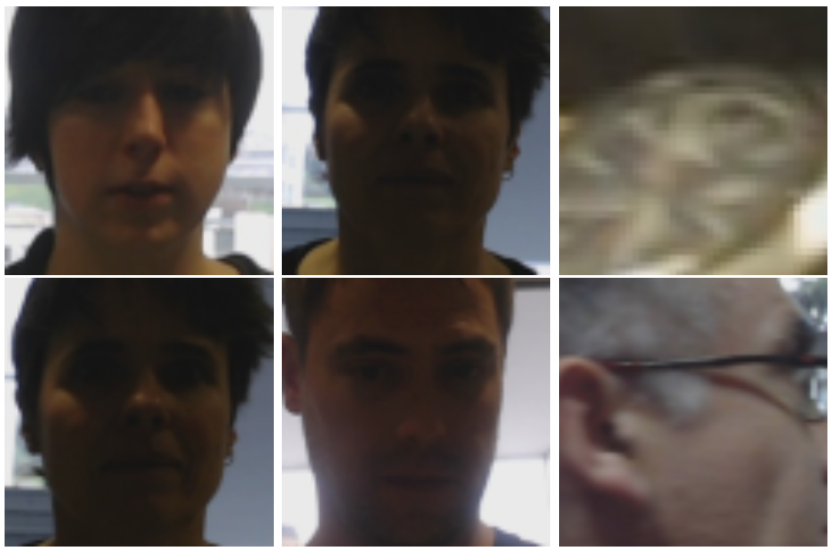


Figura 4.19: Ejemplos de errores de re-identificación. La fila superior corresponde a los probes y la fila inferior a la imagen emparejada del gallery.

		Gallery			
	Planta	**1**	**2**	**3**	**4**
Probe	**1**	98.5/98.9/98.9	92.2/92.2/92.2	77.4/77.4/77.4	80.7/81.7/81.9
	2	91.5/91.5/92.2	98.3/98.3/98.6	83.5/84.1/85.2	80.7/81.1/81.1
	3	69.3/69.3/69.7	83.5/84.4/84.4	94.8/96.5/97.8	75.8/77.1/80.2
	4	66.5/67.4/68.3	72.7/72.7/73.0	76.3/76.6/80.1	97.8/99.1/99.4

Tabla 4.11: Rank-1, Rank-5 y Rank-10 para cada planta en comparación con otras plantas.

Rank-1	**Rank-2**	**Rank-3**	**Rank-4**	**Rank-5**	**Rank-6**	**Rank-7**	**Rank-8**	**Rank-9**	**Rank-10**
81.16	86.96	88.89	91.79	93.72	93.72	96.14	96.62	96.62	97.10

Tabla 4.12: Resultados del Rank-1 al Rank-10 considerando múltiples muestras por identidad y diferentes plantas.

Una situación más frecuente y real es cuando el probe debe buscarse en las imágenes tomadas por robots de otras plantas. Para evaluar el módulo propuesto en este último escenario, nuevamente, se utilizó como probe cada imagen capturada por cada robot, y el gallery estaba compuesta por muestras capturadas por un solo robot, pero en otra planta. Teniendo en cuenta los cuatro pisos, hay 16 combinaciones de experimentos de probe-gallery. La Tabla 4.11 sintetiza los resultados obtenidos. Como era de esperar, el rendimiento más alto se obtuvo cuando el probe y el gallery correspondían al mismo piso (robot), reportando tasas de exactitud bastante optimistas, simi-

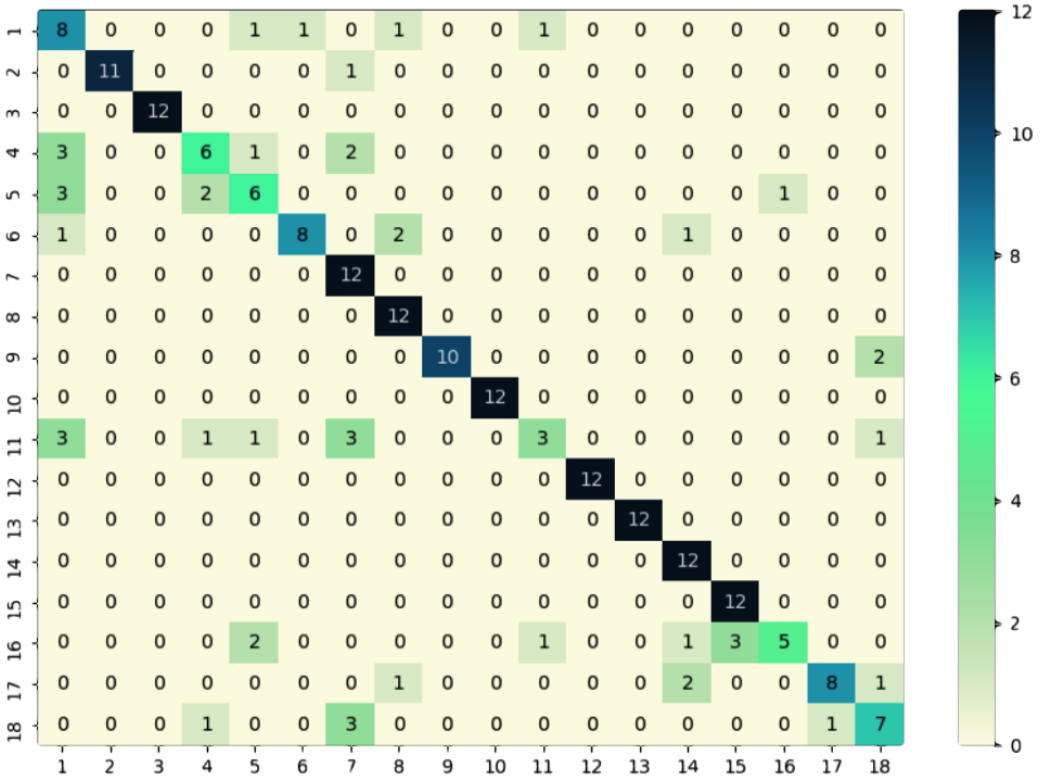


Figura 4.20: Matriz de confusión obtenida para el emparejamiento de diferentes pisos.

lares a las presentadas en la Tabla 4.10. El sistema alcanzó el 98.5 % como la puntuación máxima en términos de Rank-1 en la primera planta. En cambio, se obtuvieron resultados ligeramente inferiores en el tercer piso (94.8 % en Rank-1). En comparación con las capturas de la entrada, la luz de fondo parecía ser el factor principal que ocultaba los detalles faciales.

Al observar los resultados para este escenario, donde el probe y el gallery fueron capturadas por diferentes robots, es decir, diferentes pisos, se obtuvo un 92.2 % como Rank-1 (planta 1 frente a planta 2); y la puntuación más baja fue del 66.5 %, entre la entrada principal y la última planta. Observando la Tabla 4.9, hubo cero detecciones o tres identidades (Pisos 3 y 4). Ciertamente, no fue posible una coincidencia positiva en los experimentos cuando no había una muestra para un usuario en el gallery. En la Figura 4.19 se presentan algunos errores de emparejamiento.

Se plantea un experimento final que aprovecha la información temporal durante la interacción humano-robot. De hecho, el robot captura múltiples instancias de la cara, por lo tanto, en lugar de una sola imagen (probe), adoptamos la estrategia de emparejamiento del conjunto completo de mues-

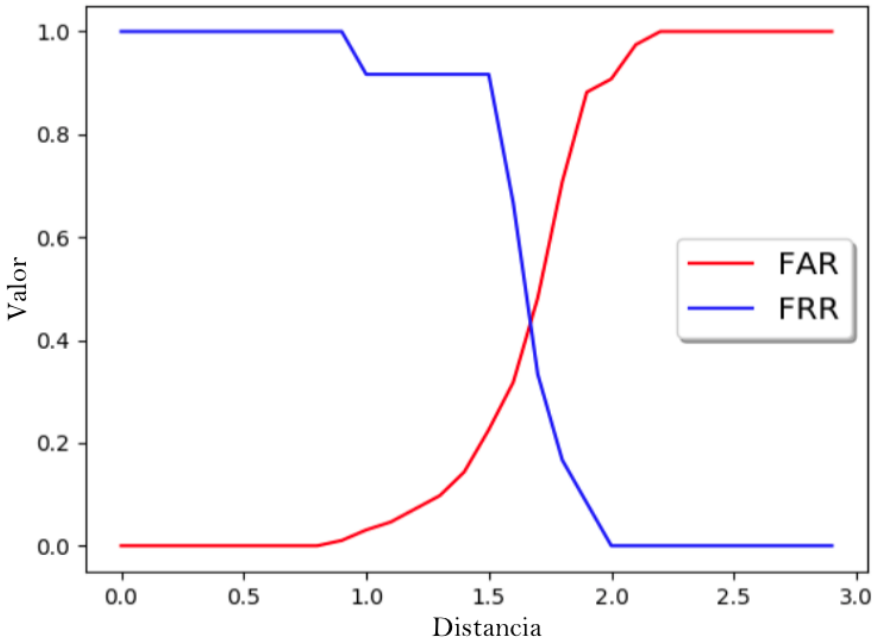


Figura 4.21: FAR y FRR obtenidos.

tras para una determinada identidad en plantas diferentes a la del probe. Para simular este escenario, se comprueba la similitud entre el individuo de la planta objetivo con las 18 identidades, escogiendo la que tenga el promedio de distancia más bajo como posible pareja. Por lo tanto, para el conjunto de n_i muestras de la identidad a en la planta de origen o, $\mathbf{x}_a^o = \left\{\mathbf{s}_{a,1}^o, \mathbf{s}_{a,2}^o, ..., \mathbf{s}_{a,n_i^o}^o\right\}$ es emparejada frente a las muestras de cualquier individuo encontrado en la planta objetivo t, $\mathbf{x}_b^t = \left\{\mathbf{s}_{b,1}^t, \mathbf{s}_{b,2}^t, ..., \mathbf{s}_{b,n_b^t}^t\right\}$, donde el objetivo se calculó obteniendo el promedio de las distancias obtenidas para cada combinación de muestra haciendo uso de las representaciones de muestra. Formalmente expresado como:

$$\mathbf{s}(\mathbf{x}_a^o, \mathbf{x}_b^t) = \frac{\sum_{i=0}^{n_a^o} \sum_{j=0}^{n_b^t} d(e(s_{a,i}^o), e(s_{b,j}^t))}{n_a^o \times n_b^t} \tag{4.15}$$

Para este enfoque, la Tabla 4.12 resume los resultados solo para los experimentos donde el origen y el piso objetivo no eran los mismos. Se puede observar que en Rank-1 se alcanzó una exactitud superior al 81 %. La matriz de confusión obtenida para las 18 identidades, considerando el Rank-1, se muestra en la Figura 4.20, evidenciando que el comportamiento fue diferente entre las identidades. Por un lado, las identidades 11 y 16 fueron mal identificadas, mientras que la 4 y la 5 se reconocían correctamente al 50 %. Por

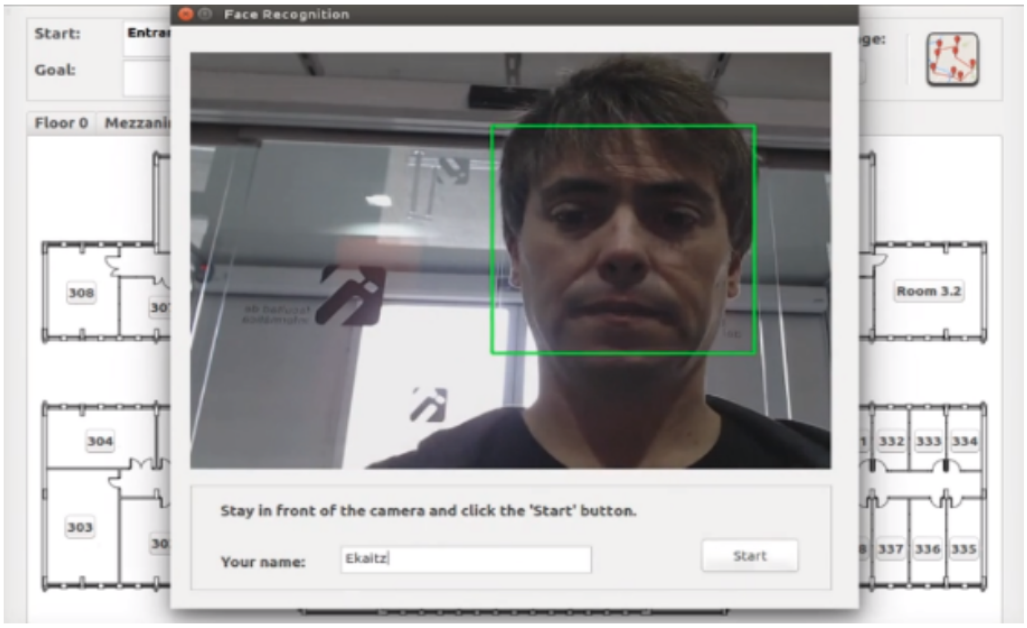


Figura 4.22: Ventana emergente que muestra la imagen del usuario capturada cuando el objetivo requiere un cambio de planta.

otro lado, ocho identidades coincidieron correctamente. Se podría argumentar que hubo un número menor de muestras para algunas de esas identidades en algunos pisos (por ejemplo, 4, 5 y 11), una circunstancia que también puede verse afectada por la calidad de la imagen. Por otro lado, es importante hacer uso de un umbral de verificación de identidad para el uso online. En la Figura 4.21, se pueden observar la medida FAR y FRR.

4.3.3.2. Integración del módulo de reconocimiento facial en el sistema Multi-Robot

Como paso final, el módulo de reconocimiento facial es adaptado e integrado en GidaBot. El módulo está vinculado al nodo de navegación multirobot y solo se activa cuando es necesario un cambio de piso. Internamente, el robot de origen debe tomar una foto del usuario y enviarla al robot ubicado en el piso de destino junto con la información relacionada con los puntos de inicio y finalización de la tarea de navegación. El segundo robot recibe el mensaje del objetivo y la imagen del visitante, asimismo, este espera al visitante que inició la tarea. Es necesario agregar una ventana emergente donde aparece la imagen a capturar (ver Figura 4.22).

Se deben definir criterios para decidir si dos imágenes corresponden a la misma persona o no. Al contrario de la fase experimental offline, el sistema

robótico implementado requiere la verificación facial haciendo uso de una sola imagen. Independientemente del método de detección de rostros, se emplea la representación de FaceNet como vector de características. Para comparar las representaciones, se define un umbral en el espacio euclídeo donde los vectores de características deben determinar si se corresponden de diferentes imágenes contenían el mismo individuo. Además, en lugar de emplear un solo método único, decidimos fusionar las capacidades de los diferentes métodos calculando la distancia entre imágenes usando los tres detectores de rostros (DLIB, VJ y MTCNN) y tomando la decisión final aplicando un umbral (TH) al mínimo de las tres distancias (*MinDistance*) obtenidas (ver Figura 4.14).

4.3.3.3. Configuración experimental online y resultados

	Exactitud	**Sensibilidad**	**Especificidad**	**F1**
MinDistance	0.95	0.95	1	0.97

Tabla 4.13: Rendimiento obtenido en el sistema multi-robot en un escenario real.

Los experimentos offline validaron el rendimiento del módulo de reconocimiento facial. Sin embargo, los resultados del mundo real pueden diferir mucho de los conjuntos de datos capturados. Por lo tanto, debe probarse la aplicabilidad y confiabilidad del enfoque en un escenario real.

Para ello, se utilizaron tres robots durante 10 días laborables durante aproximadamente 2 horas cada día, alternando mañanas y tardes, en un experimento realizado en la Facultad de Informática de la UPV/EHU de San Sebastián. Se pide a los usuarios que soliciten a uno de los robots al menos una visita guiada donde tenga que trasladarse a un piso diferente; es decir, cada tarea de guía involucró la interacción de más de un robot. En resumen, participaron cincuenta y seis usuarios diferentes (25 mujeres/31 hombres) en los experimentos.

De acuerdo con el protocolo de interacción, el sistema guía al sujeto a capturar su rostro cuando se solicita el recorrido, y posteriormente el robot del siguiente piso espera al sujeto, quien es nuevamente guiado si es verificado con la nueva captura del rostro. Por lo tanto, los impostores (o posibles falsos positivos) rara vez ocurrirían en este escenario. Sin embargo, el experimento incluyó forzar la aparición de impostores mediante la irrupción aleatoria de algunos visitantes para tratar de confundir el sistema utilizando impostores voluntarios. Después de esta observación, en la configuración experimental final, el noventa por ciento de los pares de imágenes eran genuinos en el sentido de que pertenecían a la misma persona (pero las imágenes fueron tomadas por diferentes robots). Como resultado, el diez por ciento de las veces,

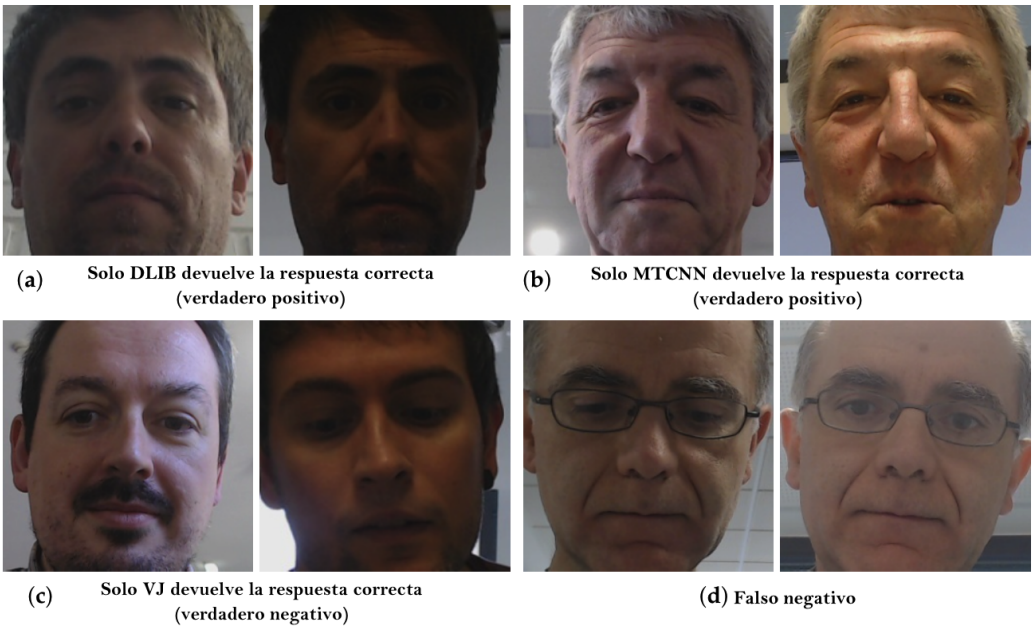


Figura 4.23: Ejemplos de casos en los que como mucho un método dio la respuesta adecuada.

Detector	**Exactitud**	***MinDistance*** **(%)**
DLIB	0.84	51.26
VJ	0.693	28.0
MTCNN	0.70	20.74

Tabla 4.14: Resultados obtenidos de cada método individualmente.

el sistema se enfrentó a imágenes que no pertenecían a la misma persona, es decir, verdaderos negativos o impostores. Todo el conjunto de datos estaba compuesto de 199 pares de imágenes que se tomaron de recorridos reales y se compararon en condiciones reales.

La Tabla 4.13 resume los resultados obtenidos para la *MinDistance* propuesta. Como se describe anteriormente, la propuesta fusionó los vectores de características con hasta tres detectores faciales, para obtener la distancia mínima. En el experimento, solo se produjeron nueve falsos negativos y no se detectó ningún falso positivo.

La fusión de tres detectores faciales evitó el sesgo presente en cada uno, proporcionando un sistema final que era menos dependiente de las fortalezas y debilidades de un solo detector, es decir, siendo más adaptativo a las diferentes condiciones que estaban presentes durante la interacción entre hombre-robot. Al analizar los datos para explorar la importancia de los di-

ID robot	Distancia recorrida (m)	Cambios de planta
1	2,313	62
2	2,757	76
3	1,973	61

Tabla 4.15: Información general de los recorridos de los robots.

ferentes detectores de rostros en el proceso, observamos que el detector de rostros DLIB proporcionó la distancia mínima en la mayoría de los casos, aproximadamente el 51 % de las veces, mientras que VJ y MTCNN tuvieron éxito respectivamente en un 28 %. y el 21 % de las veces. Estas tasas se muestran en la Tabla 4.14, incluida la tasa de acierto lograda con un único detector. Una observación rápida revela que la mejor tasa de acierto se logró utilizando el detector facial DLIB, alcanzando hasta el 84 %. En cualquier caso, era evidente que todos los detectores eran relevantes para hacer más robusto el enfoque de fusión, ya que la tasa de acierto final del método de fusión redujo el error en más de un 68 %. En la Figura 4.23 se muestran ejemplos de singularidades relacionadas con cada detector que ocurren durante los experimentos.

Alguna información adicional relacionada con la actividad de los robots durante las visitas guiadas se resume en la Tabla 4.15. Las interacciones entre los tres robots han sido equilibradas, ya que entre el número total de 199 solicitudes de guiado, interactuaron respectivamente 62, 76 y 61 veces. Sus distancias recorridas estuvieron aproximadamente entre dos y tres kilómetros, evidenciando las condiciones reales del experimento realizado.

En último lugar, hay un vídeo disponible en el canal de Youtube de RSAIT[3] se muestra a dos de los robots empleados guiando a un usuario desde el vestíbulo de entrada a una sala ubicada en el tercer piso de la facultad, además de la verificación del usuario por el segundo robot.

4.4. Diseño de base de datos en interacción hombre-máquina

4.4.1. Escenario

Para estudiar mejor el entorno de los robots, se ha decidido diseñar una base de datos debido a las carencias de los conjuntos de datos relativos a HRI. Comúnmente, los escenarios empleados para el reconocimiento de personas

[3]https://www.youtube.com/watch?v=5c4RDQg5Rsc

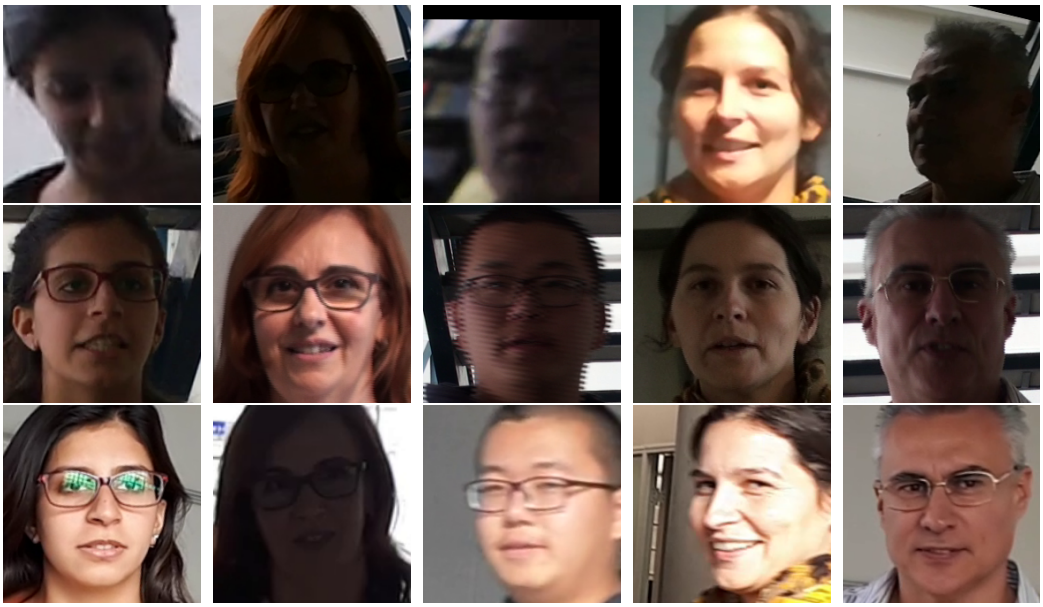


Figura 4.24: Muestras de la base de datos propuesta. Cada columna corresponde a un determinado participante y muestra una adquisición por planta. Las muestras describen variaciones en pose, iluminación, resolución y distancia de adquisición.

están situados en interiores donde las condiciones lumínicas son aproximadamente homogéneas, asimismo los dispositivos de adquisición de las imágenes. Es por ello, que en esta sección planteamos el diseño de una base de datos novedosa relacionada con la problemática de reconocimiento de personas donde además es posible aplicar técnicas de audio para su posterior diarización. Además, esta base datos proporciona un valor añadido dentro del campo de la HRI, ya que se simula el uso de robots asistentes en un escenario donde se encuentran ubicados en diferente plantas y sus características son diferentes. La recopilación de datos realizada en este trabajo implicó la adquisición de datos audiovisuales de los participantes que reproducen frases cortas frente a dispositivos de grabación en interiores. La base de datos resultante se denomina como *AveRobot* [Marras et al., 2019a].

4.4.2. Generación de la base de datos

El objetivo principal de AveRobot es proporcionar un conjunto experimental de imágenes que simule el escenario de robots de asistencia en un entorno interior semi-restrictivo, como se encuentra a menudo en edificios públicos como universidades o museos. Más precisamente, la recopilación de datos se ha llevado a cabo en el módulo de despachos del Edificio Departamental de Informática y Matemáticas de la ULPGC que es un edificio de tres

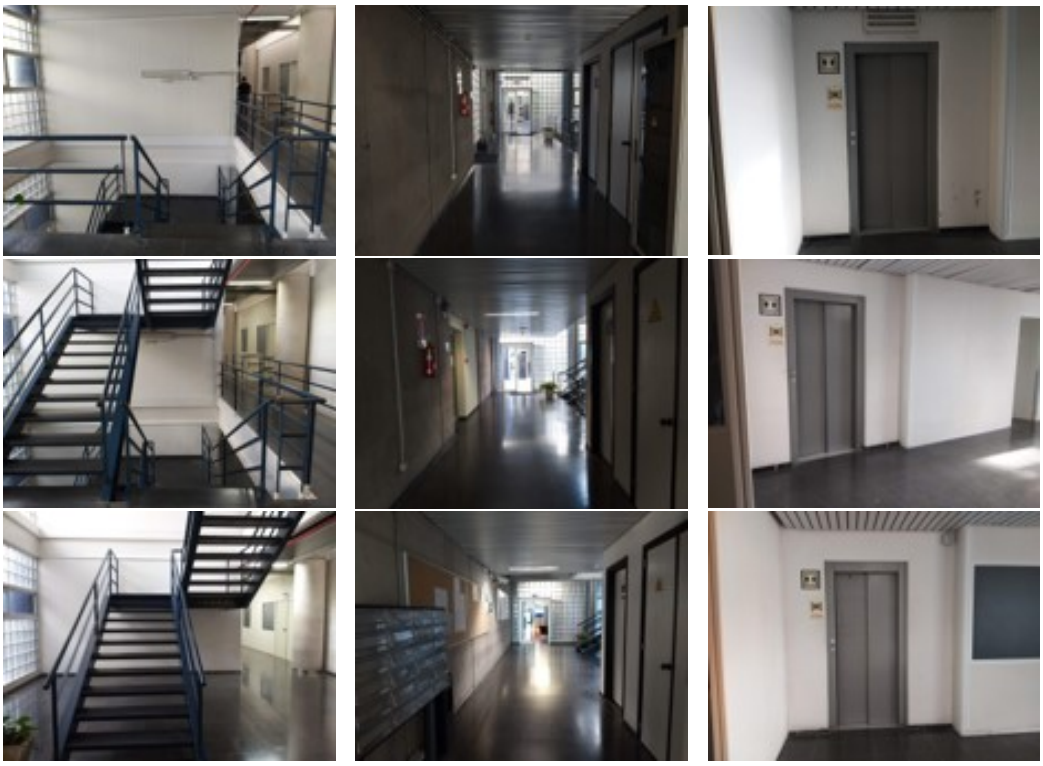


Figura 4.25: Imágenes representativas sobre cada una de las plantas y lugares en los que tiene lugar la recopilación de datos.

plantas. Teniendo en cuenta que el problema estaba relacionado con la parte sensorial del robot, no fueron necesarios robots reales, sino que se simularon mediante el uso de varias cámaras y micrófonos similares a los integrados en los robots, como además de la colocación de las cámaras a diferentes alturas. Como resultado, las interacciones en cada piso se registraron con diferentes dispositivos, simulando un total de ocho sistemas de adquisición de robots: dos en la primera planta, tres en la segunda y tres en la tercera. Además, como una persona tiene dos opciones para llegar a otro piso (es decir, usar el ascensor o las escaleras), las grabaciones se realizaron en tres lugares para cada planta: cerca de las escaleras, a lo largo del pasillo y fuera del ascensor, ver Figura 4.25. Además, en la Figura 4.24 se proporcionan muestras de rostros detectados en las grabaciones de AveRobot. Como se puede esperar, el uso de diferentes dispositivos de adquisición plantea cambios en la iluminación, geometría y resolución de la imagen y en la calidad del sonido. De hecho, la adquisición del audio incluye ruido de fondo, que consistía en conversaciones de fondo, risas, conversaciones superpuestas y acústica del lugar.

Para recopilar el conjunto de datos audiovisuales propuesto, se siguió un

procedimiento cuyos pasos se describen en los siguientes apartados.

4.4.2.1. Selección de dispositivos

ID	Modelo	Tipo	Resolución	Fps	Formato	Altura (cm)	Planta
1	Casio Exilim EXFH20	Cámara compacta	1280 × 720	30	AVI	130	0
2	Huawei P10 Lite	Cámara de móvil	1920 × 1080	30	MP4		
3	Sony HDR-XR520VE	Cámara de vídeo	1920 × 1080	30	MTS		
4	Samsung NX1000	Cámara compacta	1920 × 1080	30	MP4	120	1
5	iPhone 6S	Cámara de móvil	1920 × 1080	30	MOV		
6	Sony DCR-SR90	Cámara de vídeo	720 × 576	25	MPG		
7	Olympus VR310	Cámara compacta	1280 × 720	30	AVI	150	2
8	Samsung Galaxy A5	Cámara de móvil	1280 × 720	30	MP4		

Tabla 4.16: Especificaciones de los dispositivos de grabación usados para la construcción de la base de datos.

Para construir la base de datos se utilizaron ocho dispositivos de grabación, cada uno simulando un sistema de adquisición de un robot diferente. La Tabla 4.16 detalla sus características. Cabe señalar que los dispositivos presentan diferentes peculiaridades y son similares a los sensores integrados en robots. Las cámaras 1 y 7 tendían a generar grabaciones más borrosas. Por otro lado, las cámaras 3 y 6 grabaron vídeos usando escaneo entrelazado, a diferencia del escaneo progresivo realizado por las otras cámaras.

4.4.2.2. Configuración

Se agruparon los dispositivos por planta considerando sus diferentes características y alturas operativas. En la planta 0 se utilizaron las cámaras 1 y 2 a una altura fija de 130cm, la planta 1 incluye las cámaras 3, 4 y 5 a una altura fija de 120cm, y las cámaras 6, 7 y 8 se desplegaron en el piso 3 a una altura fija de 150cm. Por lo tanto, cada piso alberga una cámara de móvil, una cámara compacta y una cámara de vídeo, excepto el piso 0. Para asegurar que las grabaciones se hicieran en condiciones similares, se utilizan trípodes para cámaras compactas y de vídeo, mientras que las cámaras de móvil las sostiene un operador humano a la misma altura de los demás dispositivos. En la mayoría de los casos, seleccionamos una altura de grabación más baja que la de un humano porque los robots generalmente no son muy altos (por ejemplo, Pepper[4] tiene una altura de 120cm). Los dispositivos se configuraron con la resolución más alta posible a una velocidad de fotogramas constante (25fps para la Cámara 6 y 30fps para las cámaras restantes).

[4]https://www.softbankrobotics.com/emea/en/pepper

4.4.2.3. Grabación del usuario

El mismo procedimiento de grabación fue empleado para cada usuario de la base de datos. En primer lugar, para cada localización, el usuario selecciona y memoriza una frase para decirla posteriormente, esta es tomada de una lista con frases predefinidas. Mientras tanto, los dispositivos se colocaron en una posición cercana a la ubicación de destino (es decir, escaleras, pasillo y ascensor). Luego, los operadores encienden los dispositivos correspondientes al mismo tiempo, mientras el usuario se acerca a la cámara y reproduce la frase frente a los dispositivos de captura. De esta manera, en cada ubicación, se graba el mismo discurso simultáneamente con dos/tres dispositivos. El proceso se repite en cada piso y lugar seleccionando una oración diferente, y tiene una duración entre 6 y 10 minutos por usuario.

4.4.2.4. Protección de datos

Una vez finalizada la sesión, el usuario se lee y firma un acuerdo para respetar la normativa europea de protección de datos. La información brindada por el participante se limita a: su nombre completo, el número de identificación, si autoriza o no a mostrar su imagen como muestras en artículos de investigación y la firma. Se registró el sexo, altura y edad.

4.4.2.5. Etiquetado del vídeo

Los vídeos se etiquetan manualmente para realizar un seguimiento de la identidad del participante, la planta y la ubicación, la oración pronunciada y el dispositivo de grabación. Con este fin, cada vídeo se renombró mediante la siguiente convención: "UserId-FloorId-LocationId-SentenceId-DeviceId". La versión anónima se pone a disposición del público junto con el conjunto de datos.

4.4.2.6. Postprocesado del vídeo

Primero, todos los vídeos se convirtieron del formato de vídeo original al formato MP4. Luego, las caras se detectaron y alinearon con MTCNN, se redimensionaron a 224×224 píxeles y se almacenaron como imágenes en formato PNG. Los fotogramas con caras detectadas se extrajeron y guardaron, con la resolución original, como imágenes PNG en otra carpeta. Cada imagen se verificó manualmente para eliminar falsos positivos.

4.4.2.7. Postprocesado del audio

Una vez extraído el audio de cada vídeo, se almacena como un archivo con formato WAV, la parte de silencio al principio y al final del audio se eliminó mediante un proceso semiautomático usando una herramienta de segmentación de audio, Auditok[5]. Por lo tanto, los audios resultantes incluyeron solo la parte en la que el participante habla. En tercer lugar, cada audio se convirtió en secuencias de 16 bits en un solo canal a una frecuencia de muestreo de 16 kHz.

4.4.3. Estadísticas de la base de datos

El conjunto de datos propuesto contiene 2,664 vídeos de 111 participantes (65 % hombres y 35 % mujeres) que vocalizan diferentes oraciones cortas. Las frases fueron seleccionadas por el participante a partir de un conjunto predefinido de 34 oraciones diseñadas para un escenario de robots asistentes (ver Apéndice A). Las personas reunidas abarcan diferentes etnias, edades (promedio 27; desviación estándar 11; mínimo 18; máximo 60) y alturas (promedio 1.74 metros; desviación estándar 0.10 metros; mínimo 1.50 metros; máximo 1.92 metros). La Figura 4.26 muestra las distribuciones relevantes de la base de datos. El sexo, la altura y la edad de cada participante también se proporcionan junto con los vídeos. Cada persona se registró en 3 ubicaciones (es decir, escaleras, pasillo y ascensor) para cada una de las 3 plantas del edificio. Como se mencionó anteriormente, durante la toma se usaron 8 dispositivos diferentes de grabación para simular los sistemas de adquisición de robots. Los dispositivos de grabación asignados al mismo piso funcionaron simultáneamente. Por lo tanto, el conjunto de datos comprende 24 vídeos por usuario:

- 1ª planta: 2 (dispositivos) × 3 (ubicaciones) = 6 vídeos.
- 2ª planta: 3 (dispositivos) × 3 (ubicaciones) = 9 vídeos.
- 3ª planta: 3 (dispositivos) × 3 (ubicaciones) = 9 vídeos.

La duración total de los vídeos de la base de datos propuesta es de 5 horas 17 minutos, ocupando 21.8 GB. Cada participante está representado por más de 3 minutos de vídeos, cada uno con una duración media de alrededor de 7 segundos. Cabe señalar que cada vídeo incluye tres fases: (i) cuando la persona se acerca a los dispositivos, (ii) cuando habla frente a los dispositivos y (iii) cuando abandona la escena. Por lo tanto, mirando solo el contenido de

[5]https://github.com/amsehili/auditok

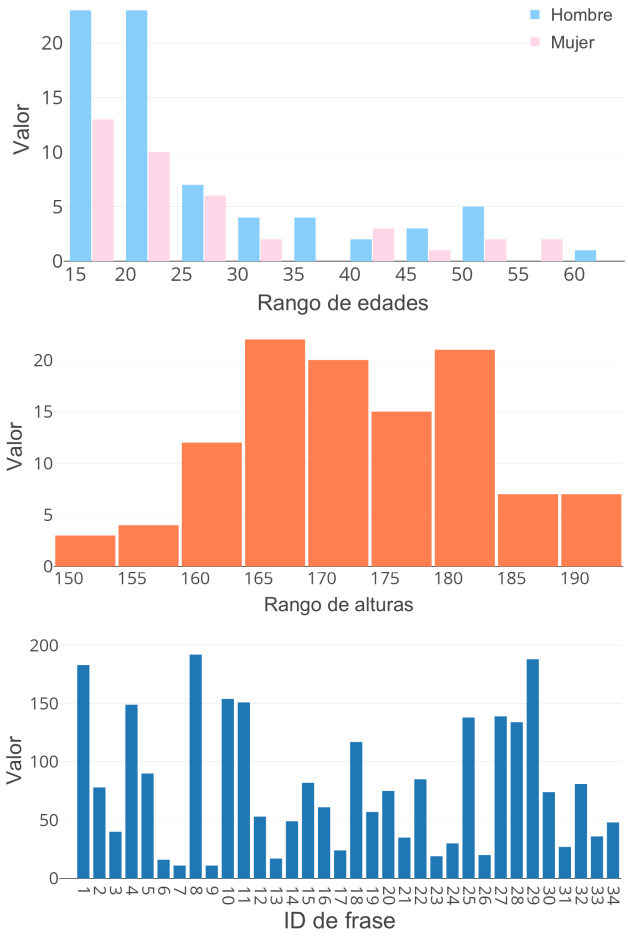


Figura 4.26: Las estadísticas de la base de datos por sexo para la distribución de edad (**superior**), distribución de la altura de los usuarios (**centro**), y la distribución de frases pronunciadas a lo largo de los vídeos (**inferior**).

la cara, cada vídeo contiene alrededor de 127 imágenes donde se ha detectado una cara, y cada usuario está representado por más de 3,000 caras detectadas. La cantidad total de caras detectadas es de 338,578. Por otro lado, mirando el contenido de voz, cada vídeo contiene alrededor de 3 segundos de discurso y cada usuario está representado por más de 1 minuto de contenido. La duración total de los datos de voz es de alrededor de 1 hora y 40 minutos.

4.5. Re-identificación multimodal

En este apartado, se describe la estrategia de fusión intermedia que conjuntamente aprende los vectores de características de la voz y de la cara,

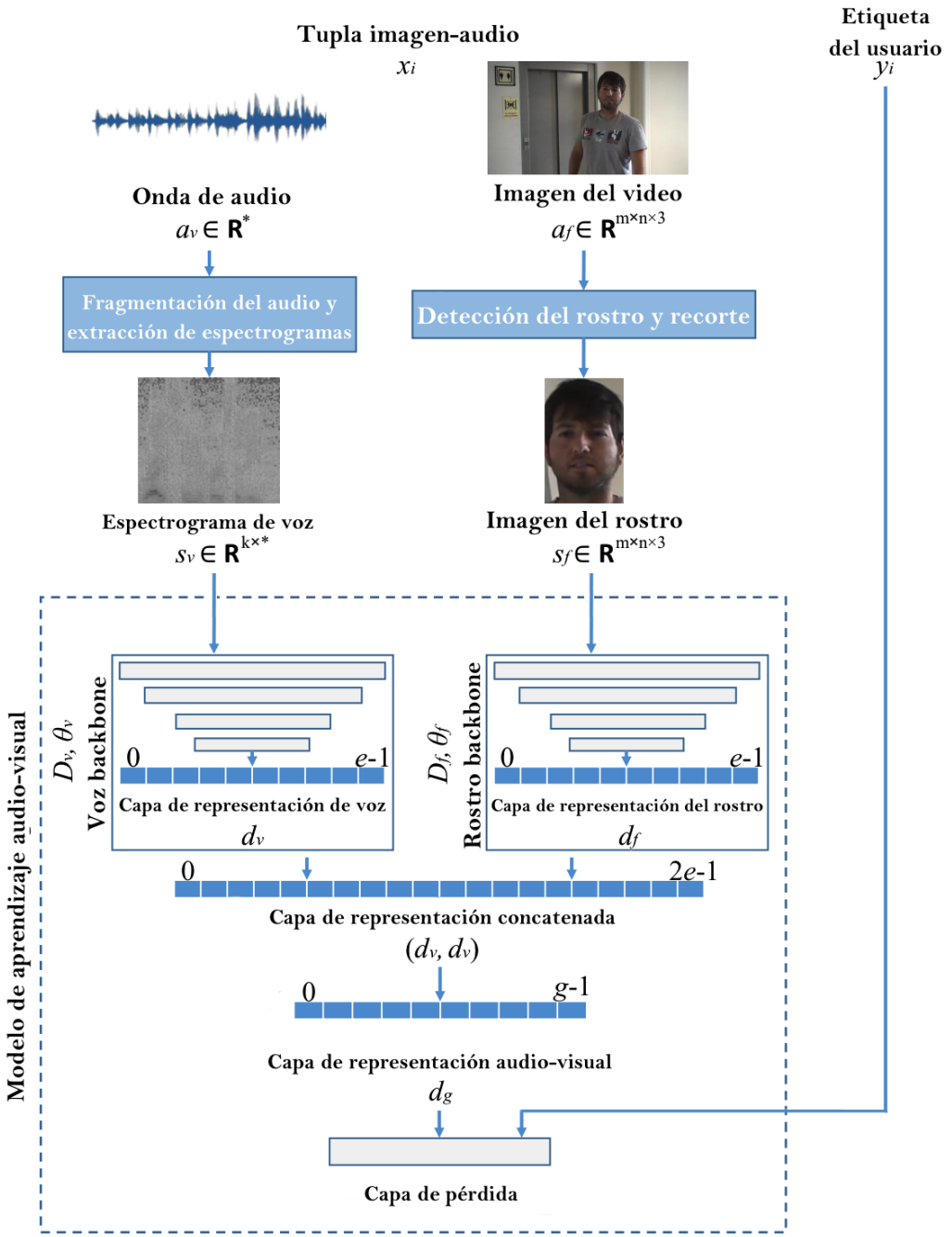


Figura 4.27: La arquitectura neuronal propuesta para la fusión multibiométrica intermedia.

incluyendo la formalización del modelo, los datos de entrada, las redes subyacentes y los detalles de entrenamiento, ver Figura 4.27.

La idea central es aprovechar las relaciones morfológicas existentes entre la biometría de la voz y la cara con el fin de investigar un entrenamiento intermodal donde cada modelo unibiométrico es apoyado por el modelo biométrico de la otra modalidad para mejorar la efectividad de sus representaciones de características. A diferencia de otros enfoques de fusión intermedia, dicha fusión multibiométrica podría ocurrir (i) en el entrenamiento para desarrollar mejores modelos unibiométricos y/o (ii) en el despliegue para aprovechar la evidencia conjunta de las dos modalidades simultáneamente.

4.5.1. Arquitectura multimodal

Denominamos $A_f \subset \mathbb{R}^{m \times n \times 3}$ el dominio de las imágenes RGB con tamaño $m \times n \times 3$. Cada imagen, $a_f \in A_f$, es preprocesada con el fin de detectar la región y los puntos claves (dos ojos, nariz y las comisuras de la boca) de la cara. Se usa una transformación afín para alinear la cara. La imagen es reescalada y cada píxel es normalizado en un rango de [0,1]. La imagen resultante, definida como $S_f \subset \mathbb{R}^{M \times N \times 3}$, es usada como entrada de la rama de la modalidad visual de nuestro modelo. En esta rama, se plantea una extracción de características explícita que genera representaciones de longitud fija en $D_f \subset \mathbb{R}^e$. Denotamos a esta etapa como $\mathcal{D}_{f_{\theta_f}} : A_f \to D_f$. Su salida corresponde al vector de características faciales.

Denominamos $A_v \subset \mathbb{R}^*$ el dominio de la señal de voz representadas digitalmente por una representación acústica visual intermedia, como un espectrograma o un banco de filtros. Cada audio $a_v \in A_v$ es convertido a un único canal. Los espectrogramas son generados empleando una ventana deslizante de Hamming, obteniendo una representación acústica s_v que corresponde al audio a_v. La normalización de la media y la varianza se realiza en cada intervalo de frecuencia del espectro. La representación resultante se utiliza como entrada de la rama de modalidad acústica de nuestro modelo. En esta rama, una extracción de características explícita que produce representaciones de longitud fija en $D_v \subset \mathbb{R}^e$. Denotamos a esta etapa como $\mathcal{D}_{v_{\theta_v}} : S_v \to D_v$. Su salida corresponde al vector de características de la voz.

Sea $D^{2 \times e}$ el dominio de los vectores de características audiovisuales generados por una concatenación simple de la representación de la cara (d_f) y la voz (d_v). Denominamos como $\mathcal{C}_\theta : (D_f, D_v) \to D^{2 \times e}$ a la etapa de concatenación de ambas modalidades aplicada después de la capa de representación de cada rama de unimodal. Luego, se aplica un paso adicional de aprendizaje del vector de características al vector concatenado $d \in D^{2 \times e}$ para obtener un vector de características único de tamaño g aprendido conjuntamente de d_f y d_v. Esta capa adicional tiene como objetivo (i) mantener independiente el tamaño del vector de características multimodal frente a los tamaños unimo-

dales, (ii) aprender representaciones más compactas y flexibles. Además, al establecer $g = e$, se pueden realizar comparaciones razonables entre representaciones dispersas unimodales y multimodales del mismo tamaño. Denotamos a esta etapa como $\mathcal{D}_{fv_{\theta_{f,v}}} : D^{2\times e} \rightarrow D^{g}$. Su salida es denominada como vector de características audiovisual.

La combinación de ambas modalidades podría generar una mejor representación del individuo y enriquecer la representación de características de una única modalidad. Esto se debe a las relaciones de la voz con el sexo y la morfología facial de las personas, por ejemplo, los hombres suelen tener un tono más bajo que las mujeres. Por lo tanto, al aprovechar la fusión, las ramas unibiométricas se ayudan entre sí para reconocer mejor a las personas. Nuestra hipótesis es que los vectores de características de cada rama deberían funcionar mejor cuando se entrenan en conjunto que cuando se entrenan por separado.

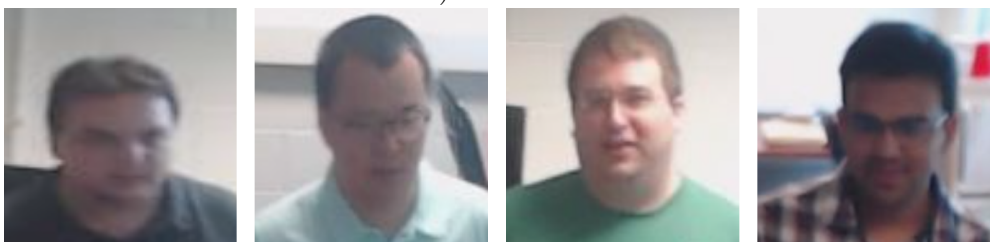
El método propuesto hace uso de arquitecturas de redes neuronales existentes, dos instancias de arquitectura de redes residuales (*ResNet-50*) se usan como extractor de vectores de características $\mathcal{D}_{f_{\theta_f}}$ y $\mathcal{D}_{v_{\theta_v}}$ dentro de las ramas de la cara y la voz, respectivamente [He et al., 2016], como se muestra en la Figura 4.27. Tal red es bien conocida por su buen rendimiento en clasificación en modalidad tanto visual como acústica [Taigman et al., 2014, Chung et al., 2018], siendo similar a una red neuronal convolucional multicapa, pero con conexiones de salto añadidas, de modo que las capas agregan residuos al mapeo de la identidad en las salidas del canal. Las capas de entrada de la arquitectura original *ResNet-50* se adaptan para la modalidad asociada a cada rama. Además, la capa completamente conectada en la parte superior de la red original se reemplaza por dos capas: una capa de aplanamiento y una capa completamente conectada cuya salida es el vector de características de la modalidad, es decir, d_f o d_v.

La estructura de fusión $\mathcal{D}_{fv_{\theta_{f,v}}}$ es instanciada mediante una capa de concatenación apilada en el modelo para combinar vectores de características de rostro y voz en el dominio $D^{2\times e}$, y una capa completamente conectada adicional donde las características importantes de la modalidad audiovisual se integran conjuntamente. La última salida representa el vector de características audiovisual previamente formalizado $d \in D^{g}$. Además, para cada capa completamente conectada, una normalización ha sido asignada antes de la función de activación para regularizar las salidas, y una capa dropout es insertada después de la activación para prevenir sobreajuste del modelo. Finalmente, la capa de salida que depende de la función de pérdida aplicada en la parte superior de la red durante el entrenamiento.

Los datos de entrenamiento están compuestos por N tuplas $\{(x_i, y_i)\}_{i=1}^{N}$

a) VoxCeleb1

b) MOBIO



c) MSU-AVIS

d) AveRobot

Figura 4.28: Muestras faciales de los conjuntos de pruebas usados para evaluar nuestro método.

donde cada muestra multibiométrica x_i corresponde a la persona asociada con la clase $y_i \in 1, ..., I$, siendo I el número de identidades diferentes representadas por N muestras. Cada muestra x_i se define como un par $x_i = (d_v, d_f)$ tal que d_v es la representación de la voz y d_f es la representación visual. Los elementos de cada par se eligen aleatoriamente entre muestras del rostro y la voz del mismo usuario; luego, se introducen secuencialmente en el modelo multibiométrico. Dicho modelo se puede integrar con cualquier función de pérdida existente. Además, se utiliza un conjunto de validación empleando

una estrategia hold-out que consta de todos los segmentos de voz y rostro de un solo vídeo seleccionado al azar por usuario.

4.5.2. Experimentos y resultados

En esta sección, se evalúa la efectividad de nuestra estrategia de fusión. Primero, detallamos los conjuntos de datos, los protocolos experimentales, los detalles de implementación y las funciones de pérdida. Luego, presentamos los resultados logrados por la estrategia de fusión sobre re-identificación y verificación, variando la función de pérdida y el conjunto de datos de prueba.

4.5.2.1. Conjuntos de entrenamiento y prueba

Consideramos los conjuntos de datos audiovisuales tradicionales para entrenar los modelos y los probamos en conjuntos de datos de diversos contextos audiovisuales (ver Figura 4.28). Esta opción permite el cálculo de puntuaciones adicionales al estado del arte en AveRobot, y permite observar cómo la estrategia afecta al rendimiento en diferentes contextos. Los conjuntos de datos audiovisuales se dividen en un conjunto de datos de entrenamiento y cuatro conjuntos de datos de prueba para replicar una configuración de conjunto de datos cruzados:

- **Conjunto de datos de entrenamiento**. *VoxCeleb1-Dev* es un conjunto de datos de identificación y verificación de intervinientes audiovisuales recopilados por [Nagrani et al., 2017] de Youtube, incluyendo 21,819 vídeos de 1,211 identidades. Es uno de los más adecuados para entrenar una red neuronal profunda debido a la amplia gama de usuarios y muestras por usuario.

- **Conjunto de prueba #1**. *VoxCeleb1-Test* es un conjunto de datos audiovisuales de identificación y verificación de intervinientes recopilados por [Nagrani et al., 2017] de Youtube, que incluye 677 vídeos de 40 identidades.

- **Conjunto de prueba #2**. *MOBIO* es un conjunto de datos de reconocimiento de rostros y intervinientes recopilados por [McCool et al., 2012] de portátiles y teléfonos móviles en un escenario controlado, que incluye 28,800 vídeos de 150 identidades.

- **Conjunto de prueba #3**. *MSU-Avis* es un conjunto de datos de reconocimiento facial y de voz recopilado por [Chowdhury et al., 2018] en escenarios de videovigilancia en interiores semicontrolados, que incluye 2,260 vídeos de 50 identidades.

- **Conjunto de prueba #4**. *AveRobot* es el conjunto de datos generado como parte de esta tesis y descrito anteriormente. Incluye 2,664 vídeos de 111 identidades.

Nos podemos dar cuenta de que la distancia de adquisición, las condiciones ambientales y la calidad de los datos varían mucho entre los conjuntos de datos, lo que los convierte en un desafío.

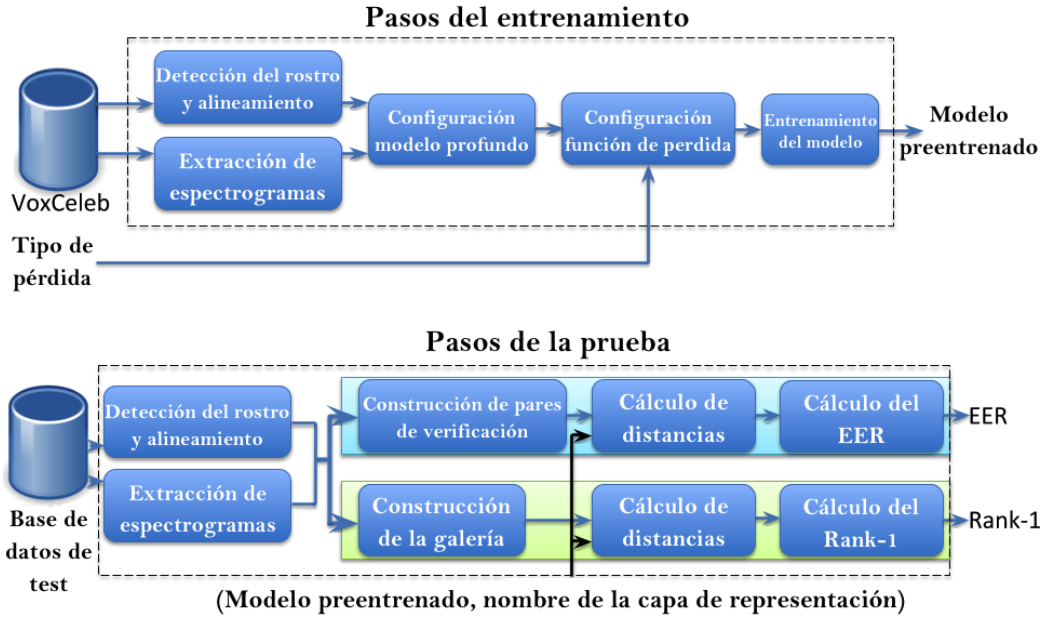


Figura 4.29: Descripción general de la evaluación experimental. Protocolos de entrenamiento y pruebas.

4.5.2.2. Configuración y protocolos de evaluación

Se plantean unos protocolos para evaluar las representaciones de las características unibiométricas y multibiométricas ante los problemas de reidentificación y verificación (Figura 4.29).

Formato de los datos. Para la rama de la cara, se analiza cada fotograma para detectar el área de la cara y los puntos de referencia a través de MTCNN. Las caras son reescaladas haciendo uso de la posición de los ojos, la nariz y el contorno de detección para generar una transformación afín de 112×112 píxeles. Con el fin de ajustarlas a la entrada de la rama y cada píxel ($[0, 255]$) en RGB se normaliza restando 127.5 y dividiéndolo entre 128.

Las imágenes resultantes se utilizan luego como entrada para la rama de la cara. Para la rama de voz, cada audio se convierte a un solo canal, flujos

de 16 bits a 16 kHz de frecuencia de muestreo para mayor coherencia. Los espectrogramas se generan luego en forma de ventana deslizante utilizando una ventana de Hamming de ancho 25 ms y un paso de 10 ms. Esto proporciona espectrogramas de tamaño 512×300 para tres segundos de audio. La normalización de la media y la varianza se realiza en cada intervalo de frecuencia del espectro. No se utiliza ningún otro preprocesamiento específico del habla. Los espectrogramas se utilizan como entrada a la rama de voz.

Representación de las características. La evaluación involucró representaciones de características unibiométricas y multibiométricas entrenadas sobre *VoxCeleb1-Dev*. Para optimizar los pesos del modelo, se entrenaron de forma independiente varias instancias de la red a través de diferentes funciones de pérdida de varias familias: *Softmax* loss [Taigman et al., 2014], *Center* loss [Wen et al., 2016], *Ring* loss [Zheng et al., 2018], y *AM-Softmax* loss [Wang et al., 2018a]. Más precisamente, para cada tipo de pérdida, se entrenaron diferentes modelos para aprender las siguientes representaciones de características:

- *Voz unimodal*, las representaciones son extraídas de d_v cuando la rama de voz se entrena independientemente.
- *Cara unimodal*, las representaciones son extraídas de d_f cuando la rama de la cara es entrenada independientemente.
- *Voz multimodal*, las representaciones extraídas de d_v cuando la rama de voz es entrenada junto con la rama de la cara.
- *Cara multimodal*, las representaciones extraídas de d_f cuando la rama de la cara es entrenada junto con la rama de voz.
- *Cara+voz multimodal*, las representaciones extraídas de d_g cuando la rama de la cara y la voz se entrenan a la vez.

Cada modelo se inicializó con pesos previamente entrenados en ImageNet. Descenso de gradiente estocástico con un declive de peso establecido en 0.0005 fue usado en mini-lotes de tamaño 512 a lo largo de 40 épocas. La tasa de aprendizaje inicial fue 0.1, y se redujo con un factor de 10 después de 20, 30 y 35 épocas.

Protocolo de re-identificación. Para cada conjunto de datos de evaluación, el protocolo tiene como objetivo evaluar cómo la representación aprendida es capaz de predecir, para un individuo de probe dado, la identidad de la persona elegida de un conjunto de identidades del gallery. Cada evaluación se realiza seleccionando al azar a 40 usuarios cada vez con el fin de (i) mantener constante el número de usuarios considerados, y (ii) mantener resultados

comparables entre los diferentes conjuntos de datos. *VoxCeleb1-Test* tiene el número mínimo de participantes entre los conjuntos de datos considerados (es decir, 40). Para cada usuario, se escoge el 80 % de los vídeos para el gallery, mientras que el restante 20 % de los vídeos como probes. Además, para cada usuario se selecciona aleatoriamente 20 imágenes/espectrogramas de los vídeos del gallery, y 100 imágenes/espectrogramas de los vídeos de probe. Luego, dado cada par imagen/espectrograma, se extrae la representación de características correspondiente. La distancia Euclídea se usa para comparar los vectores de características obtenidos de los modelos entrenados con las funciones de pérdida *Softmax*, *Center loss* y *Ring loss*, mientras que la distancia Coseno se usa para vectores de características obtenidos del modelo entrenado con la función de pérdida *AM-Softmax* debido a su diseño subyacente. Luego, se usa el Rank-1, al ser bien aceptada para evaluar las tareas de re-identificación de personas (por ejemplo, [Zheng et al., 2013]). La imagen del probe es emparejada con el conjunto de imágenes del gallery, obteniendo una lista ordenada según su similitud/distancia. El emparejamiento correcto es el correspondiente al Rank-1, para este caso particular se puede afirmar que el Rank-1 es equivalente a la tasa de acierto. Para ello, el Rank-1 se formula como la precisión en la predicción de la identidad correcta (predicción) dada la identidad espectrograma/cara conocida:

$$\textit{Rank}\text{-}1 = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.16}$$

donde TP es el verdadero positivo, TN representa el verdadero negativo, FP es el falso positivo y FN representa el falsos negativos. Finalmente, a partir de la selección de sujetos, se repitió el experimento y se promediaron los resultados.

Protocolo de verificación. Para cada conjunto de datos de prueba, el protocolo tiene como objetivo evaluar cómo las representaciones aprendidas son capaces de verificar, dado un par de caras/espectrogramas, si las caras/voces provienen de la misma persona. De cada conjunto de datos de prueba, seleccionamos aleatoriamente 40 sujetos debido a las mismas razones indicadas en el protocolo de re-identificación anterior. Luego, creamos aleatoriamente una lista de 20 vídeos (con repeticiones) para cada usuario seleccionado y, de cada uno de ellos, creamos aleatoriamente 20 pares de la misma identidad (positiva) y 20 pares de diferente identidades (negativas). Las representaciones de las características mencionadas anteriormente se consideraron como vectores de características asociados a cada cara/espectrograma. Usamos las mismas medidas de distancia aprovechadas para la re-identificación y se calcula el ERR (acrónimo en inglés de Equal Error Rate) para evaluar el rendimiento de los modelos en los pares de prueba. ERR es una métrica

de seguridad biométrica bien conocida que se usa en tareas de verificación [Jain et al., 2000]. ERR indica que la medida FAR (acrónimo en inglés de False Accept Rate) es igual a FRR (acrónimo en inglés de False Reject Rate). Ambas medidas son formuladas como:

$$\begin{aligned} FAR &= \frac{\text{número de falsos aceptados}}{\text{número de comparaciones de impostores}} \\ FRR &= \frac{\text{número de falsos rechazados}}{\text{número de comparaciones genuinas}} \end{aligned} \tag{4.17}$$

Cuanto menor sea el EER, mayor será el rendimiento del método. Por último, se promediaron los resultados.

4.5.2.3. Resultados de re-identificación

El rendimiento en los conjuntos de datos se muestra en las Figuras comprendidas entre 4.30 y 4.33. Se puede observar que los resultados varían con respecto a la modalidad, la función de pérdida de entrenamiento y el conjunto de datos. Los resultados se presentan desde estos tres puntos de vista.

Teniendo en cuenta la modalidad del rostro, las representaciones aprendidas a través de la función de pérdida Softmax aparecen como las de mejor desempeño para la configuración de rostro unimodal (Rank-1 de 0.32 a 0.74), mientras que el rendimiento de las representaciones para la configuración de la cara multimodal varía mucho entre las funciones de pérdida de entrenamiento y los conjuntos de datos de prueba. Esto significa que la estrategia de entrenamiento multibiométrico profundo se ve fuertemente afectada por la función de pérdida de entrenamiento y el conjunto de datos objetivo, mientras que las estrategias de entrenamiento unibiométricas comunes aprovechan la función de pérdida de Softmax y sus resultados solo se ven afectados por el conjunto de datos. Además, se puede observar que las representaciones faciales multimodales permiten mejorar los resultados en la re-identificación facial en escenarios desafiantes como en el conjunto de datos AveRobot. En escenarios más controlados, la fusión profunda nos permite aumentar la precisión, alcanzando resultados comparables con los obtenidos por las representaciones de características unimodales aprendidas a través de la función de pérdida Softmax en modelos faciales unibiométricos.

Se pueden realizar diferentes observaciones para la modalidad de voz. Las representaciones aprendidas a través de la función de pérdida Center loss son superiores a las representaciones aprendidas por otras funciones de pérdida en modelos de voz unimodales y multimodales. Curiosamente, las representaciones de voz multimodales funcionan peor que las representaciones de voz

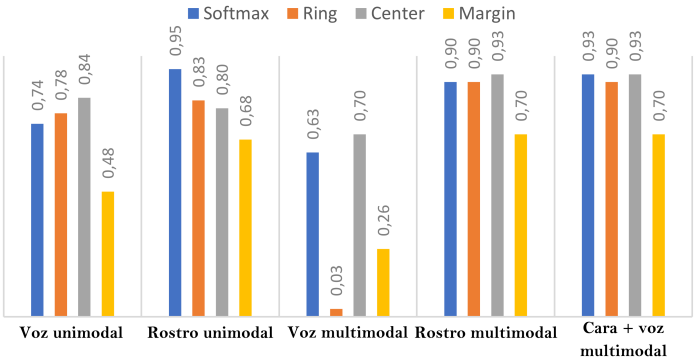


Figura 4.30: Resultados de re-identificación en VoxCeleb1-Test - Rank-1.

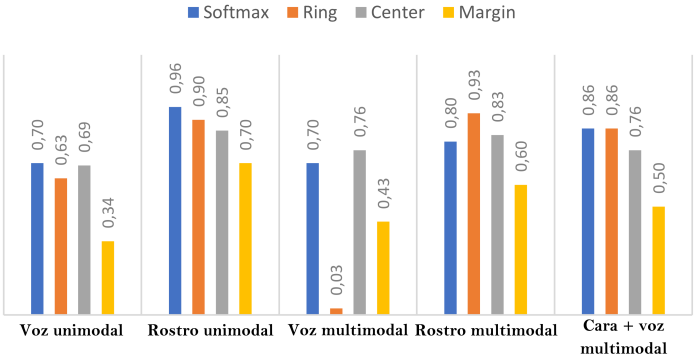


Figura 4.31: Resultados de re-identificación en MOBIO - Rank-1.

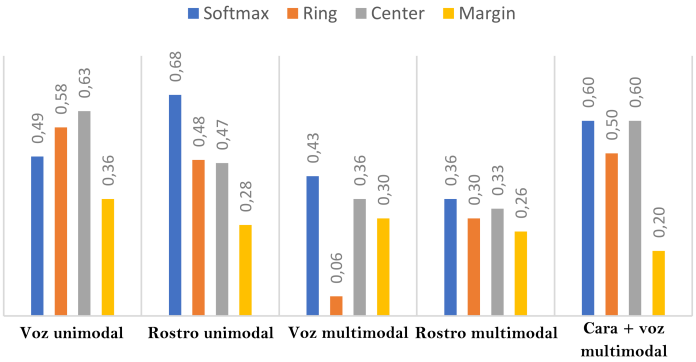


Figura 4.32: Resultados de re-identificación en MSU-Avis - Rank-1.

unimodales para cualquier función de pérdida. De ello se desprende que la biometría de voz no aprovecha en gran medida la estrategia de fusión profun-

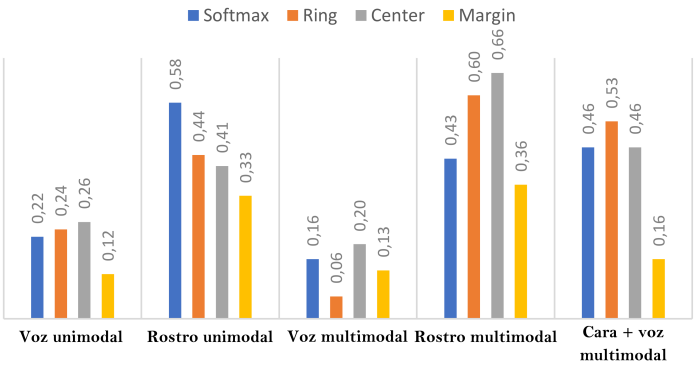


Figura 4.33: Resultados de re-identificación en Averobot - Rank-1.

da, a diferencia de lo que sucede con la biometría facial. La excepción está representada por los resultados obtenidos en las representaciones de voz multimodales en MOBIO; alcanzan resultados más altos que las representaciones de voz unimodales. Esto significa que no hay una conclusión general sobre la efectividad de las representaciones de voz multimodales, por lo que dependen del conjunto de datos. Por lo tanto, se deben realizar pruebas preliminares para seleccionar la estrategia de entrenamiento de la voz adecuada según el contexto.

En el caso de que se fusionen tanto la biometría facial como de la voz (configuración multimodal cara + voz), las representaciones aprendidas a través de las funciones de pérdida Ring loss y Center loss lograron mejores resultados que las representaciones unimodales, mientras que las representaciones aprendidas a través de las funciones de pérdida Softmax y AM-Softmax alcanzan peores resultados probablemente debido al mal desempeño de la representación de voz multimodal intermedia. De ello se deduce que la fusión del rostro y la voz durante el entrenamiento no siempre es suficiente para mejorar los resultados con respecto a las representaciones unimodales.

Entre los conjuntos de datos, VoxCeleb1-Test mostró los valores más altos de re-identificación. Esto probablemente esté relacionado con el hecho de que todos los modelos están entrenados con datos provenientes del mismo contexto, VoxCeleb1-Train. En MOBIO, las representaciones muestran resultados comparables, ya que los datos incluyen vídeos de la parte frontal grabados desde el móvil, es decir, condiciones controladas donde el reconocimiento debería ser más fácil. De manera diferente, MSU-Avis y AveRobot son los conjuntos de datos más desafiantes para obtener una buena representación. Los escenarios menos controlados son los últimos conjuntos de datos, la principal razón de que los valores de re-identificación son significativamente más bajos. En particular, el conjunto de datos AveRobot representa el escenario

más complejo, y se deben diseñar estrategias de fusión más efectivas a partir de la presentada en esta tesis.

4.5.2.4. Resultados de verificación

Las Figuras comprendidas entre 4.34 y 4.37 muestran los resultados logrados por las representaciones aprendidas en la verificación. La clasificación es ligeramente diferente con respecto a la tarea de re-identificación.

Se puede observar que las representaciones de caras multimodales logran un EER más bajo que las representaciones de caras unimodales con todos los conjuntos de datos y funciones de pérdidas. Esto significa que la fusión ayuda significativamente a crear mejores representaciones para la modalidad facial. Más precisamente, las medidas EER obtenidas mediante representaciones aprendidas a través de Ring loss y Center loss mejoran en alrededor de un 50 %, mientras que observamos una mejora de alrededor de un 25 % gracias a las representaciones aprendidas a través de las funciones de pérdida de Softmax y Margin loss. De ello se deduce que las representaciones de rostros multimodales se separan mejor entre pares genuinos e impostores.

Se obtienen resultados comparables mediante representaciones de voz multimodales, aunque la mejora con respecto a las representaciones de voz unimodales es menos evidente, es decir, entre 5 % y 10 %. Curiosamente, las representaciones de voz multimodales aprendidas a través de la función de pérdida Ring loss no obtienen buenos resultados.

Al fusionar los vectores de características de rostro y voz en una sola representación, el rendimiento de la verificación mejora en todos los conjuntos de datos, con todas las funciones de pérdida. Se puede observar una mejora de alrededor 50 % en todas las configuraciones. Las representaciones fusionadas rostro-voz funcionan bien también cuando se aprenden a través de la función de pérdida Ring loss; por lo tanto, las deficiencias experimentadas por las representaciones de voz multimodales aprendidas a través de la función de pérdida Ring loss se mitigan fusionando voz y rostro.

Los resultados de los conjuntos de datos de prueba en verificación confirman las observaciones realizadas para la tarea de re-identificación. El contexto tiene un impacto relevante en el rendimiento absoluto de los modelos, pasando de VoxCeleb1-Test a AveRobot aumentando el nivel de desafío del conjunto de datos. En particular, los resultados de verificación en los pares de AveRobot son cuatro o cinco veces peores que los obtenidos en los pares VoxCeleb1-Test. Las razones detrás de esta gran diferencia podrían estar relacionadas con las condiciones no controladas caracterizadas por caras muy oscuras y entornos muy ruidosos.

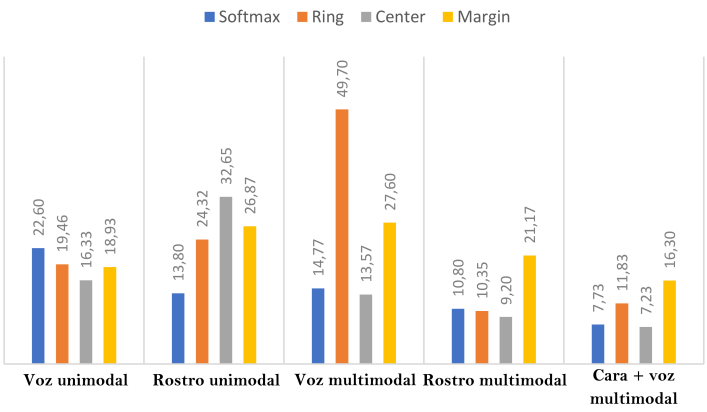


Figura 4.34: Resultados de verificación en VoxCeleb1-Test - EER.

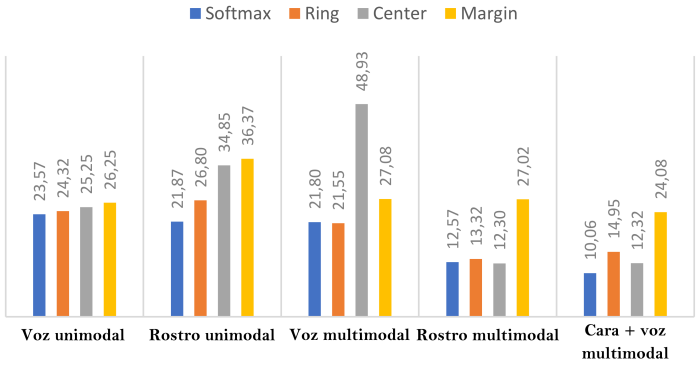


Figura 4.35: Resultados de verificación en MOBIO - EER.

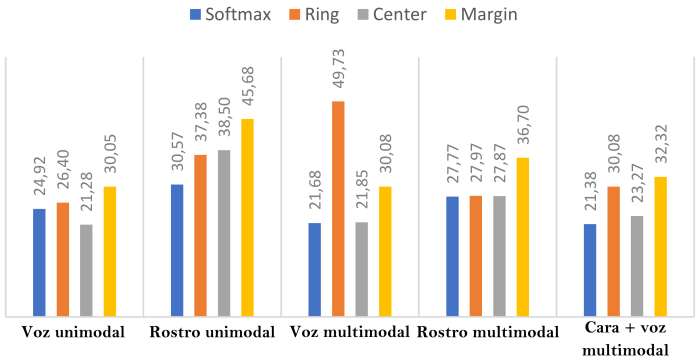


Figura 4.36: Resultados de verificación en MSU-Avis - EER.

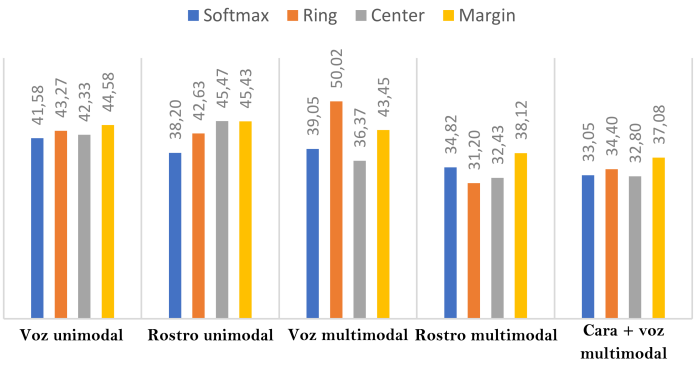


Figura 4.37: Resultados de verificación en Averobot - EER.

Capítulo 5

Conclusiones y líneas futuras

Este capítulo es una recapitulación sobre las ideas principales extraídas del trabajo desarrollado en esta tesis doctoral. Asimismo, se exponen las líneas futuras de actuación sobre el trabajo desarrollado, ya sea porque se puede mejorar lo expuesto con metodologías recientes o porque esta tesis ha abierto nuevas hipótesis de trabajo.

5.1. Conclusiones

En esta tesis doctoral se han propuesto diversas técnicas para mejorar las tareas de re-identificación de personas. Además, se ha construido un conjunto de datos lo bastante desafiante para probar diversas técnicas en un contexto realista. Por consiguiente, el diseño de esta base de datos nos abrió las puertas a desplegar un sistema de re-identifiación/verificación en robots asistentes. A continuación se desarrollan las conclusiones obtenidas:

Re-identificación de personas en debates. Se han propuesto dos estrategias, DSC y DSCK, para etiquetar a los ponentes en el contexto visual del sistema de diario en el Parlamento de Canarias. Se ha centrado en el análisis del tipo de plano, con el fin de implementar un clasificador de planos para tomar una decisión sobre el tipo proceso para la detección de personas. En el caso de planos donde aparezcan varias personas, además del interviniente, se tiene en cuenta la detección de la parte superior del cuerpo y la cara o solo la detección de la cara. Además, para evitar detecciones falsas de caras, se utilizan rasgos biométricos para la extracción de fotogramas clave. Se consigue una mejora media en términos de TRR del 4.7 % y del 9.8 % para DSC y DSCK respectivamente.

Además, se plantean cuatro estrategias para resolver los problemas que afectan a la diarización visual de intervinientes. Estas estrategias están basadas en la estructura de los planos en el vídeo. De esta forma se consiguen eliminar falsos intervinientes. Asimismo, se evalúan diferentes características extraídas de las imágenes de los intervinientes, haciendo uso de descriptores locales, diferentes patrones de imagen y medidas para comparar histogramas.

En general, el patrón HS alcanza los mejores resultados independientemente de los diferentes parámetros empleados en la configuración del experimento. También, la división de la imagen en una rejilla de 3×3 es la mejor solución. El uso de WLD es el que mejor representa el modelado del interviniente y la medida de similitud Canberra. En último lugar, pero no menos importante, el uso del plano visual más frecuente para un fragmento de audio es la estrategia con la que conseguimos reducir los errores de falsos intervinientes.

Detección de la novedad en re-identificación. Se ha presentado la re-identificación de intervinientes basada en la cara a una solución de mundo abierto para ser aplicada a los problemas de diarización en sesiones de debate parlamentario. En este escenario, es relevante la detección de la novedad de intervinientes, ya que esas identidades deben estar debidamente registradas.

Se utilizaron y evaluaron descriptores para el registro de identidad. En el problema de una clase, Resnet_T ha mostrado un buen rendimiento en la detección de novedades. El uso de HOG proporciona la mayor tasa de

exactitud para un número reducido de intervinientes. Sin embargo, cuando el número de intervinientes es mayor, WLD logra los mejores resultados. Asimismo, la mejor configuración es el Resnet_T con un clasificador SVM en la etapa de clasificación.

Los experimentos de nuestro sistema propuesto han arrojado buenos resultados con un promedio de la medida F de 71.29 % para el mejor descriptor de los vídeos. Asimismo, hemos comparado ILRA frente a las diferentes técnicas utilizadas en el reconocimiento facial en problemas de mundo cerrado, exhibiendo un incremento del 1.6 % con respecto al descriptor profundo extraído de una arquitectura basada en una red Inception Resnet empleando una tripleta.

Interacción hombre-máquina en robots asistentes. Se desarrolló e integró un módulo de re-identificación facial en un sistema de múltiples robots sociales diseñados para guiar a usuarios en escenarios de varias plantas, siendo un problema con una naturaleza compleja. Se usaron múltiples sensores y condiciones de captura e iluminación sin restricciones. En el escenario se distribuyó un solo robot por planta, requiriendo comunicación entre robots para proporcionar recorridos personalizados a los usuarios.

Se adoptó una estrategia de reconocimiento de personas, eligiendo el rostro como región de interés. El proceso de desarrollo requirió dos pasos. En primer lugar, se evaluó la adecuación de los detectores faciales actuales estándar y las incrustaciones faciales con 1,808 imágenes de prueba correspondientes a 18 identidades diferentes. Los experimentos mostraron resultados prometedores con Rank-1 superior al 81 % para la re-identificación en diferentes plantas. En segundo lugar, para implementar el reconocimiento facial en el sistema de guía turístico de múltiples robots reales, se tomó una estrategia ligeramente diferente. En lugar de utilizar un detector de rostro único, se utilizaron tres detectores de rostro simultáneamente, adoptando una propuesta de distancia mínima en el espacio de representación del rostro. Se desarrolló un experimento de la vida real de 10 días, la tasa de exactitud final alcanzó un valor superior al 95 % después de guiar a 56 usuarios en 199 interacciones. En comparación con los experimentos offline, los resultados aumentaron considerablemente. La eficiencia del módulo de reconocimiento facial fue propiciado por la fusión de las representaciones extraídas de los detectores, incrementando los resultados en un 10 % sobre la detección individual dada por MTCNN.

Incluso considerando los resultados más que satisfactorios dada la complejidad del escenario, hay algunas características del sistema que pueden afectar al sistema, como pueden ser las interrupciones en la conexión WiFi. Estas afectan la comunicación entre robots. Además, la captura de la imagen asistida por el usuario aumentó la calidad de la muestra, pero hizo que el

sistema fuera más complejo para el usuario.
Diseño de base de datos en HRI. Se propuso un flujo de procesos para recopilar datos audiovisuales en un escenario de HRI de varias plantas y la aprovechamos para crear un conjunto de datos multibiométricos que comprende modalidades de rostro y voz. Este conjunto se denomina AveRobot, diseñado para evaluar las capacidades de re-identificación y verificación de personas por robots. Incluye 111 participantes y más de 2,500 vídeos cortos.
Re-identificación multimodal. Con el fin de establecer un punto de referencia para AveRobot, se probaron diferentes técnicas para entrenar redes neuronales profundas en imágenes faciales y espectrogramas, extraídas directamente de las imágenes y los audios sin procesar, en este nuevo conjunto de datos para la re-identificación y verificación de personas. El rendimiento de este nuevo conjunto de datos se comparó frente a otros conjuntos de datos audiovisuales tradicionales. Los resultados demostraron que AveRobot es más desafiante para las técnicas empleadas debido a las condiciones incontroladas.

También, se propone una estrategia para fusionar los datos audio-visuales. una red neuronal de dos ramas alimentada con pares de caras y voces cuyo objetivo es aprender conjuntamente las representaciones de características unibiométricas y multibiométricas mediante la explotación de la correlación de características. Las ramas influyen entre sí para calcular la etiqueta de clasificación correcta después de su fusión durante el entrenamiento, de modo que la capa de representación de cada modelo unibiométrico funcione mejor que la devuelta por un modelo unibiométrico entrenado solo (una única rama). Los resultados se mejoraron aún más mediante el aprendizaje conjunto de un vector de características audiovisual.

Los modelos de rostro y voz pueden beneficiarse de la fusión intermedia. Estos son sensibles a la modalidad, función de pérdida y el contexto empleado. Los modelos faciales unibiométricos exhiben mejores resultados que los modelos de voz unibiométricos después de haber sido entrenados conjuntamente. La fusión de modalidades puede ser considerada como una solución viable para crear modelos biométricos más robustos y de confianza.

5.2. Líneas futuras

Como trabajo futuro, se plantea una serie de actividades que dan continuación a esta tesis doctoral.

En primer lugar, en lo relativo a la detección de los cambios de planos se ha realizado una implementación clásica. Actualmente hay investigaciones al respecto, donde se hacen uso de redes recurrentes para conseguir mejores

detecciones de cambios de planos. Además, puede ser beneficioso para la detección de los planos hacer uso de una estructura superior, tener en cuenta el contexto de un conjunto de planos, recurrir al concepto de escena, donde esta esta aglutinada por una serie de planos con idéntico contexto.

En segundo lugar, en lo relativo al método ILRA, como bien se ha comentado, existe una deficiencia al usar el mismo descriptor para todo el sistema. Se plantea el uso del descriptor más adecuado a la fase correspondiente. Asimismo, sería interesante hacer uso de las características del audio y de la fusión audiovisual. Además, es conveniente realizar un estudio sobre técnicas de aprendizaje profundo para remplazar la SVM que detecta la novedad.

En tercer lugar, las pruebas realizadas en robots reales abren varias vías de trabajo. Habría que continuar evaluando técnicas de reconocimiento facial. Independientemente de los modelos, cabría investigar sobre la captura e identificación automática de las imágenes con mayor calidad por parte del robot, sin la intervención humana. Por ello, es necesario agilizar los procesos y adaptar el hardware. Probar nuevos dispositivos que se integren en los robots con capacidades de computo gráfico para reducir el tráfico de la red WiFi.

En último lugar, el conjunto de datos creado, AveRobot, ha sido verificado de forma offline. El siguiente paso que corresponde es validar la propuesta en el mismo escenario pero haciendo uso de un procesamiento online, con robots reales de asistencia para dar mayor credibilidad a la base de datos propuesta. Además, se hará hincapié en el análisis de arquitecturas de redes neuronales profundas que se adapten a este contexto específico donde se fusionen diferentes tipos biométricos. Asimismo, es importante investigar en el uso de multibiométricos para poder afrontar la oclusión, mediante gafas de sol, bufandas o máscaras, de parte de la cara del usuario.

Bibliografía

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

[Anguera et al., 2012] Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370.

[Baraldi et al., 2016] Baraldi, L., Grana, C., Borghi, G., Vezzani, R., and Cucchiara, R. (2016). Shot, scene and keyframe ordering for interactive video re-use. In *11th International Conference on Computer Vision Theory and Applications*, volume 4, pages 626–631. N/A.

[Barra-Chicote et al., 2011] Barra-Chicote, R., Pardo, J. M., Ferreiros, J., and Montero, J. M. (2011). Speaker diarization based on intensity channel contribution. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4):754–761.

[Bazzani et al., 2013] Bazzani, L., Cristani, M., and Murino, V. (2013). Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144.

[Bedagkar-Gala and Shah, 2014] Bedagkar-Gala, A. and Shah, S. K. (2014). A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286.

[Birinci and Kiranyaz, 2014] Birinci, M. and Kiranyaz, S. (2014). A perceptual scheme for fully automatic video shot boundary detection. *Signal Processing: Image Communication*, 29(3):410–423.

[Boucenna et al., 2016] Boucenna, S., Cohen, D., Meltzoff, A. N., Gaussier, P., and Chetouani, M. (2016). Robots learn to recognize individuals from

imitative encounters with people and avatars. *Scientific Reports*, 6. In press.

[Bourdev et al., 2011] Bourdev, L., Maji, S., and Malik, J. (2011). Describing people: A poselet-based approach to attribute classification. In *International Conference on Computer Vision*, pages 1543 – 1550.

[Bredin and Gelly, 2016] Bredin, H. and Gelly, G. (2016). Improving speaker diarization of tv series using talking-face detection and clustering. In *ACM International Conference on Multimedia*, pages 157–161. ACM.

[Castrillón et al., 2011] Castrillón, M., Déniz, O., Hernández, D., and Lorenzo, J. (2011). A comparison of face and facial feature detectors based on the Viola–Jones general object detection framework. *Machine Vision and Applications*, 22(3):481–494.

[Castrillón-Santana et al., 2013] Castrillón-Santana, M., Lorenzo-Navarro, J., and Ramón-Balmaseda, E. (2013). Improving gender classification accuracy in the wild. In *18th Iberoamerican Congress on Pattern Recognition*, pages 270–277.

[Castrillón-Santana et al., 2017] Castrillón-Santana, M., Lorenzo-Navarro, J., and Ramón-Balmaseda, E. (2017). Multi-scale score level fusion of local descriptors for gender classification in the wild. *Multimedia Tools and Applications*, 76(4):4695–4711.

[Chai et al., 2013] Chai, Z., Sun, Z., Tan, T., and Mendez-Vazquez, H. (2013). Local salient patterns—a novel local descriptor for face recognition. In *2013 International Conference on Biometrics (ICB)*, pages 1–6. IEEE.

[Chan-Lang et al., 2017] Chan-Lang, S., Pham, Q.-C., and Achard, C. (2017). Closed and open-world person re-identification and verification. In *International Conference on Digital Image Computing: Techniques and Applications*, pages 1–8.

[Chandola et al., 2009] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *Computing Surveys*, 41(3):1–58.

[Chen et al., 2012] Chen, H., Gallagher, A., and Girod, B. (2012). Describing clothing by semantic attributes. In *European Conference on Computer Vision (ECCV)*, pages 609–623.

[Chen et al., 2010] Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., and Gao, W. (2010). WLD: A robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1705–1720.

[Chowdhury et al., 2018] Chowdhury, A., Atoum, Y., Tran, L., Liu, X., and Ross, A. (2018). Msu-avis dataset: Fusing face and voice modalities for biometric recognition in indoor surveillance videos. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3567–3573.

[Chung et al., 2018] Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

[Cielniak and Duckett, 2003] Cielniak, G. and Duckett, T. (2003). Person identification by mobile robots in indoor environments. In *1st International Workshop on Robotic Sensing*, pages 5–pp.

[Clark and Boswell, 1991] Clark, P. and Boswell, R. (1991). Rule induction with cn2: Some recent improvements. In Kodratoff, Y., editor, *Machine Learning — EWSL-91*, pages 151–163, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Clifton et al., 2013] Clifton, D. A., Clifton, L., Hugueny, S., Wong, D., and Tarassenko, L. (2013). An extreme function theory for novelty detection. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):28–37.

[Cong et al., 2010] Cong, D.-N. T., Khoudour, L., Achard, C., Meurie, C., and Lezoray, O. (2010). People re-identification by spectral classification of silhouettes. *Signal Processing*, 90(8):2362–2374.

[Correa et al., 2012] Correa, M., Hermosilla, G., Verschae, R., and Ruiz-del Solar, J. (2012). Human detection and identification by robots using thermal and visual information in domestic environments. *Journal of Intelligent and Robotic Systems*, 66(1-2):223–243.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[Cosar et al., 2017] Cosar, S., Coppola, C., Bellotto, N., et al. (2017). Volume-based human re-identification with RGB-D cameras. In *VISIGRAPP (4: VISAPP)*, pages 389–397.

[Cover, 1999] Cover, T. M. (1999). *Elements of Information Theory*. John Wiley & Sons.

[Cruz et al., 2008] Cruz, C., Sucar, L. E., and Morales, E. F. (2008). Real-time face recognition for human-robot interaction. In *8th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–6.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In Schmid, C., Soatto, S., and Tomasi, C., editors, *International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 886–893.

[Deng et al., 2018] Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., and Jiao, J. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 994–1003.

[Dinca and Hancke, 2017] Dinca, L. M. and Hancke, G. P. (2017). The fall of one, the rise of many: a survey on multi-biometric fusion methods. *IEEE Access*, 5:6247–6289.

[Ding et al., 2017] Ding, H., Zhou, S. K., and Chellappa, R. (2017). Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 118–126.

[Ding et al., 2020] Ding, Y., Xu, Y., Zhang, S.-X., Cong, Y., and Wang, L. (2020). Self-supervised learning for audio-visual speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4367–4371. IEEE.

[Dollar et al., 2010] Dollar, P., Belongie, S., and Perona, P. (2010). *The Fastest Pedestrian Detector in the West*. BMVA Press.

[Dollar et al., 2011] Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761.

[Egozcue et al., 2003] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

[El Khoury et al., 2014] El Khoury, E., Sénac, C., and Joly, P. (2014). Audiovisual diarization of people in video content. *Multimedia Tools and Applications*, 68(3):747–775.

[Enzweiler and Gavrila, 2008] Enzweiler, M. and Gavrila, D. M. (2008). Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195.

[Fenu and Marras, 2018] Fenu, G. and Marras, M. (2018). Controlling user access to cloud-connected mobile applications by means of biometrics. *IEEE Cloud Computing*, 5(4):47–57.

[Fenu et al., 2018] Fenu, G., Marras, M., and Boratto, L. (2018). A multi-biometric system for continuous student authentication in e-learning platforms. *Pattern Recognition Letters*, 113:83–92.

[Freire-Obregón et al., 2014] Freire-Obregón, D., Castrillón-Santana, M., Lorenzo-Navarro, J., and Ramón-Balmaseda, E. (2014). Automatic clothes segmentation for soft biometrics. In *Proceedings of IEEE International Conference on Image Processing*, pages 4972–4976.

[Freire-Obregón et al., 2021] Freire-Obregón, D., Rosales-Santana, K., Marín-Reyes, P. A., Penate-Sanchez, A., Lorenzo-Navarro, J., and Castrillón-Santana, M. (2021). Improving user verification in human-robot interaction from audio or image inputs through sample quality assessment. *Pattern Recognition Letters*, 149:179–184.

[Friedland et al., 2009] Friedland, G., Hung, H., and Yeo, C. (2009). Multimodal speaker diarization of real-world meetings using compressed-domain video features. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4069–4072.

[Galiyawala et al., 2018] Galiyawala, H., Shah, K., Gajjar, V., and Raval, M. S. (2018). Person retrieval in surveillance video using height, color and gender. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE.

[Gallagher and Chen, 2008] Gallagher, A. and Chen, T. (2008). Clothing cosegmentation for recognizing people. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[Gao and Ma, 2014] Gao, G. and Ma, H. (2014). To accelerate shot boundary detection by reducing detection region and scope. *Multimedia Tools and Applications*, 71(3):1749–1770.

[García-Olalla et al., 2014] García-Olalla, O., Alegre, E., Fernández-Robles, L., and González-Castro, V. (2014). Local oriented statistics information booster (losib) for texture classification. In *2014 22nd international conference on pattern recognition*, pages 1114–1119. IEEE.

[Ge et al., 2019] Ge, Y., Zhang, R., Wang, X., Tang, X., and Luo, P. (2019). Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5337–5345.

[Gebru et al., 2017] Gebru, I., Ba, S., Li, X., and Horaud, R. (2017). Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39. In press.

[Gheissari et al., 2006] Gheissari, N., Sebastian, T. B., and Hartley, R. (2006). Person reidentification using spatiotemporal appearance. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1528–1535.

[Golla and Sharma, 2018] Golla, M. R. and Sharma, P. (2018). Performance evaluation of facenet on low resolution face images. In *International Conference on Communication, Networks and Computing*, pages 317–325. Springer.

[Gonzalez-Sosa et al., 2018] Gonzalez-Sosa, E., Fierrez, J., Vera-Rodriguez, R., and Alonso-Fernandez, F. (2018). Facial soft biometrics for recognition in the wild: Recent works, annotation, and cots evaluation. *IEEE Transactions on Information Forensics and Security*, 13(8):2001–2014.

[Guo et al., 2016] Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Irigoien and Arenas, 2008] Irigoien, I. and Arenas, C. (2008). Inca: New statistic for estimating the number of clusters and identifying atypical units. *Statistics in Medicine*, 27(15):2948–2973.

[Jain et al., 2000] Jain, A., Hong, L., and Pankanti, S. (2000). Biometric identification. *Communications of the ACM*, 43(2):90–98.

[Jain et al., 2004] Jain, A. K., Dass, S. C., and Nandakumar., K. (2004). Soft biometric traits for personal recognition systems. In *International Conference on Biometric Authentication*, pages 731–738.

[Jun and Kim, 2012] Jun, B. and Kim, D. (2012). Robust face detection using local gradient patterns and evidence accumulation. *Pattern Recognition*, 45(9):3304–3316. Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011).

[Kalantidis et al., 2013] Kalantidis, Y., Kennedy, L., and Li, L. (2013). Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos. In *International Conference on Multimedia Retrieval (ICMR)*.

[Kazemi and Sullivan, 2014] Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874.

[Kim et al., 2010] Kim, H., Chung, W., and Yoo, Y. (2010). Detection and tracking of human legs for a mobile service robot. In *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pages 812–817.

[King, 2009] King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758.

[Knerr et al., 1990] Knerr, S., Personnaz, L., and Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. In Soulié, F. F. and Hérault, J., editors, *Neurocomputing*, pages 41–50, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Koide and Miura, 2016] Koide, K. and Miura, J. (2016). Identification of a specific person using color, height, and gait features for a person following robot. *Robotics and Autonomous Systems*, 84:76–87.

[Kumar et al., 2011] Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2011). Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1962–1977.

[Le et al., 2015] Le, N., Wu, D., Meignier, S., and Odobez, J.-M. (2015). Eumssi team at the mediaeval person discovery challenge. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, number EPFL-CONF-213706.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the Institute of Electrical and Electronics Engineers*, 86(11):2278–2324.

[Li et al., 2012] Li, B., Lian, X.-C., and Lu, B.-L. (2012). Gender classification by combining clothing, hair and facial component classifiers. *Neurocomputing*, 76(1):18–27.

[Li et al., 2018] Li, X., Wu, A., and Zheng, W.-S. (2018). Adversarial open-world person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 280–296.

[Liao et al., 2015] Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206.

[Lin et al., 2019] Lin, Y.-R., Su, W.-H., Lin, C.-H., Wu, B.-F., Lin, C.-H., Yang, H.-Y., and Chen, M.-Y. (2019). Clothing recommendation system based on visual information analytics. In *2019 International Automatic Control Conference (CACS)*, pages 1–6. IEEE.

[Liu et al., 2017] Liu, H., Hu, L., and Ma, L. (2017). Online RGB-D person re-identification based on metric model update. *CAAI Transactions on Intelligence Technology*, 2(1):48–55.

[Liu et al., 2012] Liu, L., Zhao, L., Long, Y., Kuang, G., and Fieguth, P. (2012). Extended local binary patterns for texture classification. *Image and Vision Computing*, 30(2):86–99.

[Liu et al., 2018] Liu, T., Ye, X., and Sun, B. (2018). Clothing and carrying invariant gait-based gender recognition. In *2018 International Conference on Image and Video Processing, and Artificial Intelligence*, volume 10836, page 108360X. International Society for Optics and Photonics.

[Lorenzo-Navarro et al., 2018] Lorenzo-Navarro, J., Castrillón-Santana, M., Gómez, M., Herrera, A., and Marín-Reyes, P. A. (2018). Automatic counting and classification of microplastic particles. In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods*.

[Lorenzo-Navarro et al., 2013] Lorenzo-Navarro, J., Castrillón-Santana, M., and Hernández-Sosa, D. (2013). On the use of simple geometric descriptors provided by RGB-D sensors for re-identification. *Sensors*, 13(7):8222–8238.

[Lorenzo-Navarro et al., 2014] Lorenzo-Navarro, J., Castrillón-Santana, M., Ramón-Balmaseda, E., and Freire-Obregón, D. (2014). Evaluation of lbp and hog descriptors for clothing attribute description. In *Video Analytics*

for Audience Measurement - First International Workshop (VAAM). Revised Selected Papers., volume 8811 of *Lecture Notes in Computer Science*, pages 53–65. Springer.

[Mäkinen and Raisamo, 2008] Mäkinen, E. and Raisamo, R. (2008). Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):541–547.

[Marín-Reyes et al., 2019] Marín-Reyes, P. A., Irigoien, I., Sierra, B., Lorenzo-Navarro, J., Castrillón-Santana, M., and Arenas, C. (2019). ILRA: Novelty detection in face-based intervener re-identification. *Symmetry*, 11(9):1154.

[Marín-Reyes et al., 2016] Marín-Reyes, P. A., Lorenzo-Navarro, J., Castrillón-Santana, M., and Sánchez-Nielsen, E. (2016). Shot classification and keyframe detection for vision based speakers diarization in parliamentary debates. In *Conference of the Spanish Association for Artificial Intelligence*, pages 48–57. Springer.

[Marín-Reyes et al., 2017] Marín-Reyes, P. A., Lorenzo-Navarro, J., Castrillón-Santana, M., and Sánchez-Nielsen, E. (2017). Who is really talking? a visual-based speaker diarization strategy. In *International Conference on Computer Aided Systems Theory*, pages 322–329. Springer.

[Marín-Reyes et al., 2018] Marín-Reyes, P. A., Palazzi, A., Bergamini, L., Calderara, S., Lorenzo-Navarro, J., and Cucchiara, R. (2018). Unsupervised vehicle re-identification using triplet networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 166–171.

[Markou and Singh, 2003] Markou, M. and Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497.

[Markov and Nakamura, 2008] Markov, K. and Nakamura, S. (2008). Improved novelty detection for online gmm based speaker diarization. In *Annual Conference of the International Speech Communication Association*, pages 363–366.

[Marras et al., 2019a] Marras, M., Marín-Reyes, P. A., Lorenzo-Navarro, J., Castrillón-Santana, M., and Fenu, G. (2019a). AveRobot: An audio-visual

dataset for people re-identification and verification in human-robot interaction. In *International Conference on Pattern Recognition Applications and Methods*.

[Marras et al., 2019b] Marras, M., Marín-Reyes, P. A., Lorenzo-Navarro, J., Castrillón-Santana, M., and Fenu, G. (2019b). Deep multi-biometric fusion for audio-visual user re-identification and verification. In *International Conference on Pattern Recognition Applications and Methods*, pages 136–157. Springer.

[Martinson and Lawson, 2011] Martinson, E. and Lawson, W. (2011). Learning speaker recognition models through human-robot interaction. In *IEEE International Conference on Robotics and Automation*, pages 3915–3920.

[Martinson et al., 2013] Martinson, E., Lawson, W., and Trafton, J. G. (2013). Identifying people with soft-biometrics at fleet week. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 49–56.

[McCool et al., 2012] McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matejka, P., Cernockỳ, J., Poh, N., Kittler, J., Larcher, A., Levy, C., et al. (2012). Bi-modal person recognition on a mobile phone: using mobile phone data. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 635–640.

[Moghaddam and Yang, 2002] Moghaddam, B. and Yang, M.-H. (2002). Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):707–711.

[Nagrani et al., 2017] Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *ArXiv Preprint arXiv:1706.08612*.

[Nguyen et al., 2016] Nguyen, D. T., Li, W., and Ogunbona, P. O. (2016). Human detection from images and videos: A survey. *Pattern Recognition*, 51:148–175.

[Ojala et al., 1994] Ojala, T., Pietikainen, M., and Harwood, D. (1994). Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *International Conference on Pattern Recognition*, volume 1, pages 582–585.

[Ojala et al., 2002] Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987.

[Ojansivu and Heikkilä, 2008] Ojansivu, V. and Heikkilä, J. (2008). Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing*, pages 236–243. Springer.

[Ortega-León et al., 2019] Ortega-León, C., Marín-Reyes, P. A., Lorenzo-Navarro, J., Castrillón-Santana, M., and Sánchez-Nielsen, E. (2019). Video categorisation mimicking text mining. In *International Work-Conference on Artificial Neural Networks*, pages 292–301. Springer.

[Ouellet et al., 2014] Ouellet, S., Grondin, F., Leconte, F., and Michaud, F. (2014). Multimodal biometric identification system for mobile robots combining human metrology to face recognition and speaker identification. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 323–328.

[Parra et al., 2019] Parra, O., Rodriguez, I., Jauregi, E., Lazkano, E., and Ruiz, T. (2019). Gidabot: A system of heterogeneous robots collaborating as guides in multi-floor environments. *Intelligent Service Robotics*, 12(4):319–332.

[Pimentel et al., 2014] Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.

[Prosser et al., 2010] Prosser, B. J., Zheng, W.-S., Gong, S., Xiang, T., and Mary, Q. (2010). Person re-identification by support vector ranking. In *The British Machine Vision Conference*, volume 2, page 6.

[Ristani et al., 2016] Ristani, E., Solera, F., Zou, R. S., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. *ArXiv*, abs/1609.01775.

[Rodriguez et al., 2020] Rodriguez, I., Zabala, U., Marín-Reyes, P. A., Jauregi, E., Lorenzo-Navarro, J., Lazkano, E., and Castrillón-Santana, M. (2020). Personal guides: Heterogeneous robots sharing personal tours in multi-floor environments. *Sensors*, 20(9):2480.

[Roth et al., 2014] Roth, P. M., Hirzer, M., Köstinger, M., Beleznai, C., and Bischof, H. (2014). Mahalanobis distance learning for person re-identification. In *Person Re-identification*, pages 247–267. Springer.

[Sánchez-Nielsen et al., 2019] Sánchez-Nielsen, E., Chávez-Gutiérrez, F., and Lorenzo-Navarro, J. (2019). A semantic parliamentary multimedia approach for retrieval of video clips with content understanding. *Multimedia Systems*, 25(4):337–354.

[Sánchez-Nielsen et al., 2017] Sánchez-Nielsen, E., Chávez-Gutiérrez, F., Lorenzo-Navarro, J., and Castrillón-Santana, M. (2017). A multimedia system to produce and deliver video fragments on demand on parliamentary websites. *Multimedia Tools and Applications*, 76(5):6281–6307.

[Schroff et al., 2015] Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

[Sermanet et al., 2013] Sermanet, P., Kavukcuoglu, K., Chintala, S., and LeCun, Y. (2013). Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633.

[Shaoshuai et al., 2020] Shaoshuai, W., Lili, G., Ling, C., Weiyong, L., Yong, C., Jingyi, Z., and Ling, F. (2020). A case report of neonatal covid-19 infection in china. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*.

[Shi and Jain, 2019] Shi, Y. and Jain, A. K. (2019). Probabilistic face embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6902–6911.

[Shneiderman, 1996] Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343.

[Singh and Aggarwal, 2015] Singh, R. D. and Aggarwal, N. (2015). Novel research in the field of shot boundary detection–a survey. In *Advances in Intelligent Informatics*, pages 457–469. Springer.

[Sinha et al., 2013] Sinha, A., Chakravarty, K., and Bhowmick, B. (2013). Person identification using skeleton information from kinect. In *Proc. Intl. Conf. on Advances in Computer-Human Interactions*, pages 101–108.

[Sinha et al., 2006] Sinha, P., Balas, B., Ostrovsky, Y., and Russell, R. (2006). Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962.

[Stolcke and Yoshioka, 2019] Stolcke, A. and Yoshioka, T. (2019). Dover: A method for combining diarization outputs. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 757–763. IEEE.

[Sujatha and Mudenagudi, 2011] Sujatha, C. and Mudenagudi, U. (2011). A study on keyframe extraction methods for video summary. In *2011 International Conference on Computational Intelligence and Communication Networks*, pages 73–77.

[Szegedy et al., 2017] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI Conference on Artificial Intelligence*.

[Taigman et al., 2014] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708.

[Tan and Triggs, 2010] Tan, X. and Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, 19(6):1635–1650.

[Tao et al., 2016] Tao, D., Guo, Y., Song, M., Li, Y., Yu, Z., and Tang, Y. Y. (2016). Person re-identification by dual-regularized kiss metric learning. *IEEE Transactions on Image Processing*, 25(6):2726–2738.

[Tranter and Reynolds, 2006] Tranter, S. E. and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565.

[Tsipas et al., 2020] Tsipas, N., Vrysis, L., Konstantoudakis, K., and Dimoulas, C. (2020). Semi-supervised audio-driven tv-news speaker diarization using deep neural embeddings. *The Journal of the Acoustical Society of America*, 148(6):3751–3761.

[Ustinova et al., 2017] Ustinova, E., Ganin, Y., and Lempitsky, V. (2017). Multi-region bilinear convolutional neural networks for person re-identification. In *14th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–6.

[Vezzani et al., 2013] Vezzani, R., Baltieri, D., and Cucchiara, R. (2013). People reidentification in surveillance and forensics: A survey. *Computing Surveys*, 46(2):1–37.

[Viola and Jones, 2004] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

[Viola et al., 2003] Viola, P., Jones, M. J., and Snow, D. (2003). Detecting pedestrians using patterns of motion and appearance. In *Proc. of the International Conference on Computer Vision*, volume 2, pages 734–741.

[Wang et al., 2018a] Wang, F., Cheng, J., Liu, W., and Liu, H. (2018a). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

[Wang, 2013] Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1):3–19.

[Wang et al., 2018b] Wang, Y., Shen, J., Petridis, S., and Pantic, M. (2018b). A real-time and unsupervised face re-identification system for human-robot interaction. *Pattern Recognition Letters*.

[Wen et al., 2016] Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer.

[Wong et al., 1995] Wong, C., Kortenkamp, D., and Speich, M. (1995). A mobile robot that recognizes people. In *Proceedings of 7th IEEE international conference on tools with artificial intelligence*, pages 346–353.

[Xia et al., 2011] Xia, L., Chen, C.-C., and Aggarwal, J. K. (2011). Human detection using depth information by kinect. In *International Workshop on Human Activity Understanding from 3D Data in conjunction with Computer Vision and Pattern Recognition*.

[Yamaguchi et al., 2012] Yamaguchi, K., Kiapour, M. H., Ortiz, L. E., and Berg, T. L. (2012). Parsing clothing in fashion photographs. In *Computer Vision and Pattern Recognition*.

[Yang and Yu, 2011] Yang, M. and Yu, K. (2011). Real-time clothing recognition in surveillance videos. In *18th IEEE International Conference on Image Processing*, pages 2937–2940.

[Ye et al., 2021] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Yong et al., 2012] Yong, S.-P., Deng, J. D., and Purvis, M. K. (2012). Novelty detection in wildlife scenes through semantic context modelling. *Pattern Recognition*, 45(9):3439–3450.

[Yu et al., 2017] Yu, H.-X., Wu, A., and Zheng, W.-S. (2017). Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 994–1002.

[Yu et al., 2020] Yu, H.-X., Wu, A., and Zheng, W.-S. (2020). Unsupervised person re-identification by deep asymmetric metric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):956–973.

[Zhang et al., 2020] Zhang, H., Sun, Y., Liu, L., Wang, X., Li, L., and Liu, W. (2020). Clothingout: a category-supervised gan model for clothing segmentation and retrieval. *Neural computing and applications*, 32(9):4519–4530.

[Zhang et al., 2016] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

[Zheng et al., 2015] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[Zheng et al., 2012] Zheng, W.-S., Gong, S., and Xiang, T. (2012). Transfer re-identification: From person to set-based verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2650–2657.

[Zheng et al., 2013] Zheng, W.-S., Gong, S., and Xiang, T. (2013). Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):653–668.

[Zheng et al., 2018] Zheng, Y., Pal, D. K., and Savvides, M. (2018). Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5097.

[Zheng et al., 2017] Zheng, Z., Zheng, L., and Yang, Y. (2017). A discriminatively learned cnn embedding for person reidentification. *Transactions on Multimedia Computing, Communications, and Applications*, 14(1):1–20.

[Zhong et al., 2018] Zhong, Z., Zheng, L., Li, S., and Yang, Y. (2018). Generalizing a person retrieval model hetero- and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232.

[Zhu et al., 2018] Zhu, X., Wu, B., Huang, D., and Zheng, W.-S. (2018). Fast open-world person re-identification. *IEEE Transactions on Image Processing*, 27(5):2286–2300.

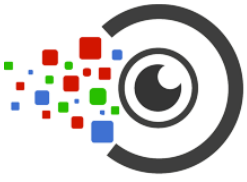
Apéndice A

Frases

01. Can you help me, please?
02. Can you repeat it slower, please?
03. Could I know the room where the Programming exam is?
04. Could you repeat that please?
05. Could you spell that for me, please?
06. Could you tell me the room where the Erasmus meeting is?
07. Do you mind saying me where Marco's office is?
08. Good morning, I am Mike.
09. How do we get to Prof. Santana's office from here?
10. How much time does it take?
11. I'm afraid I didn't get that.
12. I'm sorry, I didn't catch the room number.
13. I am trying to get in touch with Prof. Carlos.
14. I just wanted to ask, where is the toilet?.
15. I need to do some photocopying.
16. I would like to meet Prof. Castro.
17. I would like to speak with Prof. Marin.
18. Is Prof. Fernandez there?
19. Is Prof. Gonzales available?
20. Is there a free room where I can study?
21. May I please speak to Prof. Sanchez?
22. What floor is the secretariat?
23. What floor is Prof. Valverde's office?
24. When is a good time to meet Prof. Castro?
25. Where are the stairs?
26. Where is Prof. Garcia's office?
27. Where is the canteen?
28. Where is the Erasmus office?
29. Where's the lift?

30. Where's the photocopier?
31. Where is Prof. Navarro
32. Where is the toilet?
33. What's the reason for this recording?

SIANI



AImage Lab

University of Modena and Reggio Emilia

UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA