Universidad Autónoma de Madrid
Escuela politécnica superior

Trabajo Fin de Máster

INCORPORACIÓN DE ATRIBUTOS FACIALES A SISTEMAS DE RECONOCIMIENTO FACIAL

Máster universitario en Investigación e Innovación en Tecnologías de la Información y las Comunicaciones

Autora: BLÁZQUEZ CORTÉS, Marta

Tutor: VERA RODRIGUEZ, Ruben
Departamento de Tecnología Electrónica y de las Comunicaciones

Febrero 2019

INCORPORACIÓN DE ATRIBUTOS FACIALES A SISTEMAS DE RECONOCIMIENTO FACIAL

Autora: BLÁZQUEZ CORTÉS, Marta

Tutor: VERA RODRIGUEZ, Ruben
Ponente: FIERREZ AGUILAR, Julian

BiDA Lab

Biometrics and Data Pattern Analytics

Departamento de Tecnología Electrónica y de las Comunicaciones

Febrero 2019

Resumen

Resumen

La aparición de redes neuronales profundas ha provocado un gran progreso en el ámbito de la biometría. Los sistemas de reconocimiento facial son cada vez más utilizados y cada vez requieren una mayor precisión. Un modo habitual de mejorar estos sistemas es el refuerzo mediante atributos característicos de cada persona, los llamados soft biometrics. El género, la edad o la raza son algunos de los atributos más habituales.

Al analizar el rendimiento de los sistemas de reconocimiento facial se observan diferencias dentro de cada grupo demográfico. Atendiendo al género, las mujeres son las que peores resultados obtienen. Para el caso de la raza, son las personas de raza negra o asiática las que normalmente presentan más dificultades en el reconocimiento facial. Este problema radica principalmente en los conjuntos de entrenamiento con los que los modelos han aprendido. Estos no suelen estar balanceados y se refleja en los resultados cuando analizamos cada clase. Normalmente las bases de datos incluyen más hombres y más identidades de raza blanca.

En este trabajo se desarrollan sistemas específicos para los grupos demográficos de género y raza. Los resultados experimentales demuestran que utilizando modelos entrenados con imágenes pertenecientes a una única clase se mejora el rendimiento de un sistema de reconocimiento facial genérico que ha sido entrenado con imágenes de todas las clases.

Se proponen también dos estimadores para los atributos de género y raza. Se compara el rendimiento del sistema cuando la información de dichos atributos es obtenida de manera manual, es decir mediante etiquetas y cuando se extrae de manera automática.

Además se propone un sistema más completo que fusiona la información de género y raza. Y se analizan las alternativas de fusión a nivel de features y a nivel de scores.

Palabras Clave

Reconocimiento Facial, Reconocimiento Biométrico, Redes Neuronales Convolucionales, Triplet Loss, Género, Raza, Soft Biometrics.

Abstract

The research in deep neural networks has produced a great improvement in the world of biometrics. Facial recognition systems are used more often and require a higher accuracy. A common way of improving these systems is the reinforcement through characteristic attributes from each person which are known as soft biometrics. The gender, age or ethnic group are the most common attributes.

Analyzing the performance of facial recognition systems, differences are observed within each demographic group. Considering the gender, women obtain the worst results. Regarding the ethnicity group, dark skin persons or asian have more difficulties in the facial recognition. This problem is mainly due to the training sets used for the learning process of the models. These are not usually balanced and that is reflected in the results obtained for each class. Usually datasets include more men and more white race identities.

In this project, specific models are developed for the demographic groups of gender and ethnicity. The experimental results show that using trained models with images from a single class, it is possible to improve the performance of a generic facial recognition system trained with images from all classes.

Two estimators for the gender and ethnic group attributes are also proposed. System performance is compared when race and gender information is obtained automatically or manually, through label.

Moreover, a more complete system is proposed combining gender and ethnic group information. Proposing a fusion of this information at the scores or the features level.

Key words

Face Recognition, Biometric Recognition, Convolutional Neural Network, Triplet Loss, Gender, Race, Soft Biometrics.

Agradecimientos

A mi abuelo Rufo.

Índice general

Índice de Figuras

Índice de Tablas

1 Introducción

1.1. Motivación

En los últimos años, los avances en el reconocimiento facial son cada vez mayores. Con la aparición de las redes neuronales profundas, el progreso es cada vez más rápido y las alternativas se disparan. El desarrollo de *DeepFace* [27] en 2014, supuso una revolución en el ámbito del reconocimiento facial. Este sistema cuenta una amplia base de datos, como es *Facebook*. Esto supone una gran ventaja a la hora del entrenamiento de las redes. Tanto Facebook como Google mantienen privadas las bases de datos con las que entrenan sus modelos. Por esta razón surge la necesidad de crear amplias bases de datos y arquitecturas para obtener resultados competentes.

Incluso sistemas de reconocimiento facial desarrollados por grandes compañías como *Google* tienen errores. En 2015 el algoritmo de *Google Photos* cometió el inapropiado error clasificando a una pareja de raza negra en la categoría de *gorilas*. Posteriormente, el error fue subsanado eliminando directamente la categoría de gorilas [1]. Lo que plantea la idea de que determinadas razas presentan más dificultades para los sistemas de reconocimiento facial. Esta es una de las motivaciones de este trabajo. Estudiar las dificultades que presentan las distintas clases de los grupos demográficos de género y raza para sistemas de reconocimiento facial.

Muchas de las bases de datos actuales con las que se entrenan los modelos de reconocimiento facial no están balanceadas, atendiendo al género y la raza. En la mayoría de los casos el número de hombres es superior al de mujeres, y el número de personas de raza blanca es superior al del resto de razas. Esta es una de las razones por las que el rendimiento de los sistemas no es el mismo para cada grupo demográfico. Las mujeres y las personas de raza negra son las clases que peores resultados obtienen [14], [6]. En el caso de las mujeres hay artículos que atribuyen los malos resultados al uso de maquillaje, lo que hace que aumente la variabilidad *intra-class*.

En [14] se hace hincapié en la importancia de los conjuntos de entrenamiento. Se demuestra que un entrenamiento con imágenes de una sola clase puede mejorar el rendimiento de un sistema general. Por ejemplo, las mujeres obtendrán mejores resultados con un sistema entrenado únicamente con mujeres que utilizando que uno entrenado con hombres y mujeres.

Esta es la motivación principal de este trabajo. Conseguir mejorar el rendimiento de un sistema de reconocimiento facial, entrenándolo para cada una de las clases de los grupos demográficos de género (femenino y masculino) y raza (asiática, blanca, negra). Estos han sido

los grupos demográficos estudiados porque según [7], junto con la edad, son los atributos más discriminatorios.

1.2. Objetivos

El principal objetivo que persigue este trabajo es la mejora de un sistema de reconocimiento facial que ha sido entrenado con una base de datos no balanceada mediante el entrenamiento de modelos para cada clase de los grupos demográficos de género (masculino y femenino) y raza (asiática, blanca y negra).

Para lograr este propósito, se marcan los siguientes objetivos específicos:

- Analizar la complejidad que presenta cada clase de los grupos demográficos de género y raza en el reconocimiento facial.
- Desarrollo de los modelos específicos para cada una de las clases, partiendo de los modelos pre-entrenados del estado del arte *VGGFace* y *Resnet50*.
- Desarrollo de dos estimadores automáticos de los atributos de género y raza. Se comparará el rendimiento del sistema completo cuando la información de género y raza se obtiene de manera automática o mediante etiquetas manuales.
- Implementación de un sistema completo, el cual aprovecha la información tanto de género como de raza para mejorar el reconocimiento facial. Se proponen dos métodos de fusión, uno de ellos a nivel de features y otro a nivel de scores.

1.3. Estructura de la memoria

Este trabajo está organizado de la siguiente manera:

- *Capítulo 1*: Plantea la motivación, los objetivos y la estructura de este trabajo.
- *Capítulo 2*: Resume el estado del arte. Ofrece una introducción a la biometría y a los soft biometrics así como un resumen de artículos de interés relacionados con este trabajo.
- *Capítulo 3*: Describe las bases de datos utilizadas en este trabajo.
- *Capítulo 4*: Describe los métodos propuestos para la realización de este trabajo.
- *Capítulo 5*: Detalla los experimentos llevados a cabo, así como los resultados obtenidos en cada uno de ellos.
- *Capítulo 6*: Expone las conclusiones obtenidas y el trabajo futuro.

2

Trabajos relacionados. Estado del arte

En este capítulo, se introducen brevemente algunos conceptos comunes de los sistemas biométricos. Centrándose principalmente en los soft biometrics y sistemas de reconocimiento facial. Se resumen algunos de los trabajos más relevantes y que son de especial interés para la realización de este trabajo.

En concreto se hablará de los distintos usos que se les puede dar a los soft biometrics. Se comentarán distintos modos de extracción automática de atributos faciales. Para terminar, se comentará un trabajo que analiza el impacto de algunos soft biometrics demográficos en el reconocimiento facial.

2.1. Introducción a la biometría

Un sistema de reconocimiento biométrico es aquel que hace uso de rasgos propios de las personas a la hora de su autenticación. Cada vez es mayor el uso de estos sistemas en nuestro día a día. Dos de las razones fundamentales son la comodidad que ofrecen, ya que no es necesario un objeto, como una llave o una tarjeta, o la necesidad de recordar una contraseña, a la hora de autenticarse. La otra de las razones es la seguridad que estos sistemas garantizan, lo que hace que sean bien acogidos por la mayoría de la población.

Para garantizar la comodidad y seguridad de los sistemas biométricos, es necesario que los rasgos característicos, ya sean físicos (iris, huella dactilar, cara, etc...) o basados en el comportamiento (voz, firma, tecleo, paso, etc...), cumplan una serie de requisitos [12].

- *Universalidad*, toda persona debe tener ese rasgo biométrico.
- *Unicidad*, el rasgo tiene que ser único de cada persona, para poder distinguir entre dos sujetos.
- *Permanencia*, que el rasgo se mantenga invariable en el tiempo. O que al menos sea estable a largo plazo.
- *Coleccionable*, es necesario que el rasgo biométrico se pueda medir cuantitativamente.

Además, los sistemas biométricos deben tener en cuenta aspectos prácticos como la relación calidad precio, la robustez ante ataques y la aceptabilidad por parte del usuario. Todos ellos teniendo en cuenta la aplicación final para la que el sistema biométrico se va a utilizar.

Un sistema biométrico puede funcionar de dos modos. Hablamos de *identificación* cuando el sistema realiza una comparación de uno frente a todos los usuarios. Estos usuarios han sido previamente registrados en la base de datos. La respuesta del sistema será la lista de usuarios puntuados según el parecido a la muestra a identificar. La primera identidad de la lista la que más probabilidades tenga de ser el usuario reclamado. Mientras que en el modo de *verificación* se realiza una comparación entre dos usuarios. La respuesta del sistema será si ambos usuarios son o no la misma persona, dependiendo de si la comparación supera o no un umbral determinado. Como es de esperar los sistemas basados en identificación requieren mayor coste computacional.

Muchos de los sistemas biométricos actuales requieren de instrumentos complejos a la hora de adquirir los datos con los que posteriormente se realiza el reconocimiento. Esto hace que en ocasiones sean sistemas poco manejables o de alto coste. Como puede ser el caso de los sensores para adquirir imágenes del iris, o algunos sensores para la adquisición de huellas dactilares. No es el caso del reconocimiento facial, uno de los rasgos biométricos más utilizados. Es un método no intrusivo y apenas requiere colaboración con el usuario. En aplicaciones como por ejemplo el desbloqueo de dispositivos móviles, donde el entorno está controlado, los resultados son favorables. Sin embargo, en aplicaciones de videovigilancia suelen aparecer problemas debido a variaciones de iluminación o de pose, oclusiones, poca resolución en las imágenes, etc. En estos caso los resultados no son tan buenos. Los avances en el desarrollo de redes neuronales profundas, ha reducido las tasas de error en los sistemas de reconocimiento facial, incluso en escenarios complejos [23]. Además se buscan otras técnicas que mejoren el rendimiento de los sistemas de reconocimiento facial, como puede ser el apoyo de estos sistemas mediante los llamados soft biometrics.

2.2. Soft Biometrics

Existen múltiples maneras de identificar a una persona mediante rasgos biométricos, como pueden ser la firma, el iris, la huella dactilar, la cara. Sin embargo, algunas de estas características biométricas resultan complejas de obtener y en ocasiones se requiere el uso de cámaras o sensores especiales para recoger la información y posteriormente identificar a la persona. Existen otros tipos de características que resultan más sencillas a la hora de describir a una persona, como pueden ser la edad, el género, la altura, la raza, etc. Estos rasgos los podemos denominar características blandas, en inglés *soft biometrics*.

En 1892, Alphonse Bertillon propuso un sistema basado en soft biometrics, utilizando características como medidas fisiológicas, la complexión o marcas de cicatrices o tatuajes, para la identificación de criminales. Sin embargo, este sistema carecía de automatización, era difícil de adminstrar y no garantizaba la variación entre individuos [12].

Una de las ventajas que ofrecen los soft biometrics es que estos pueden ser extraídos a cierta distancia, sin la necesidad de que el usuario esté en contacto con el dispositivo. Además, no es necesario trabajar con imágenes de alta calidad, como pueden ser necesarios para sistemas biométricos como el reconocimiento de iris o de huella dactilar. Es por eso que resultan útiles en aplicaciones de videovigilancia, donde el escenario suele ser más complejo y aparecen inconvenientes como oclusiones, variaciones de pose o de iluminación. Otra ventaja que ofrecen los soft biometrics es la semejanza con la interpretación que puede realizar una persona, es decir, los soft biometrics resultan fácilmente descriptibles y entendibles para el ser humano. En escenarios de videovigilancia o forenses, cuando se describe verbalmente al sospechoso puede ser de gran ayuda.

Sin embargo, los soft biometrics resultan insuficientes por sí mismos a la hora de reconocer a un usuario. Es por eso que en la mayoría de los casos sirven como apoyo secundario a sistemas biométricos primarios. Normalmente con sistemas de reconocimiento facial, debido a su similitud, ya que la mayoría de soft biometrics son extraídos de características faciales. Aunque también existen trabajos que estudian la combinación de soft biometrics con otros sistemas biométricos primarios, como por ejemplo la estimación de género a partir de la huella dactilar, la firma o la voz [4]

Los sistemas que utilizan soft biometrics, se pueden clasificar en tres modalidades, atendiendo a la manera en la que hagan uso de los soft biometrics (ver figura 2.1): I) Utilizando los soft biometrics para el reconocimiento directamente *(Bag of Soft Biometrics)*, II) Reduciendo el campo de búsqueda al utilizar otro sistema biométrico primario, III) Fusionando soft biometrics con sistemas biométricos primarios (*(hard biometrics*).

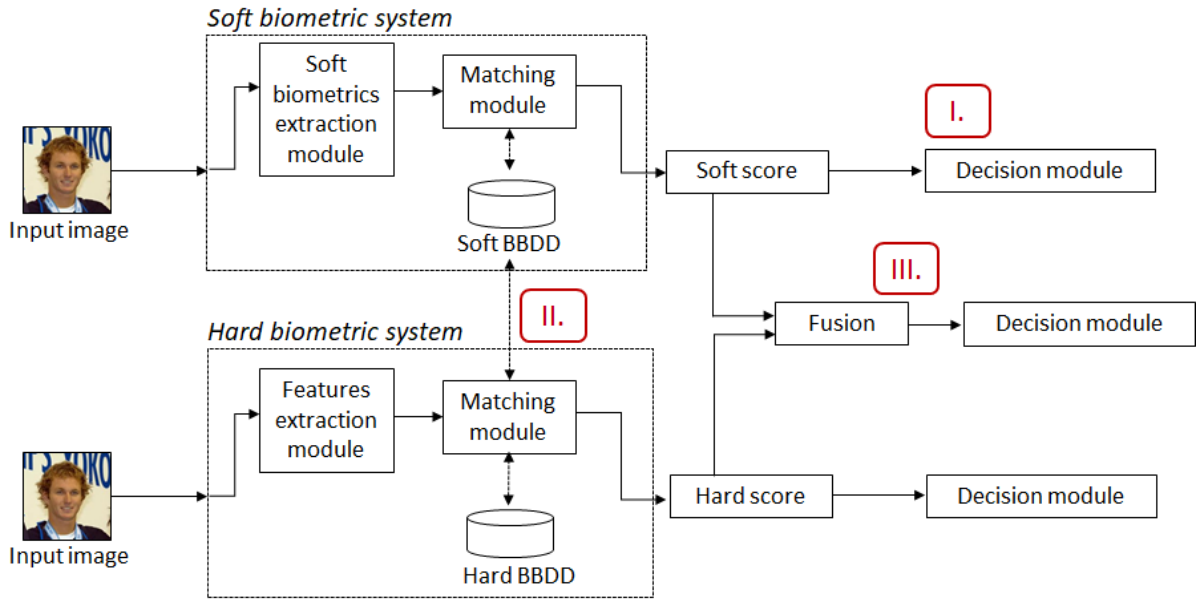

Figura 2.1: Esquema de las diferentes modalidades con las que se pueden utilizar los soft biometrics.

I.) Bag of Soft Biometrics

Trabajos previos proponen el reconocimineto de personas basado únicamente en soft biometrics. Estos pueden ser extraídos de manera automática o manual, y pueden ser únicamente atributos faciales o utilizar también atributos corporales. Por ejemplo, en [16] se clasifican 73 atributos faciales y la sonrisa. Para ello se divide la cara en 9 regiones, para entrenar los distintos atributos, para finalmente clasificarlos mediante SVM. Los resultados obtenidos aseguran entre un 80 % y un 90 % de precisión sobre la base de datos *LFW* [10].

II.) Reducción del campo de búsqueda

Otra de las maneras de hacer uso de los soft biometrics es realizando un primer filtrado basado en soft biometrics, para posteriormente proceder al reconocimiento mediante un sistema de hard biometric. De este modo el campo de búsqueda se reduce, por tanto, el número de comparaciones a la ahora del reconocimiento es también menor, y como consecuencia el tiempo necesario para realizar esta tarea también disminuye [5]. La rapidez en el tiempo de búsqueda es la principal ventaja, sin embargo, hay que tener en cuenta que este modo de utilizar los soft biometrics es sensible a los posibles fallos cometidos en la estimación de los soft biometrics. Corremos el riesgo

de dejar fuera del campo de búsqueda al sospechoso si la estimación del soft biometric no es precisa, y por tanto reducir el rendimiento del sistema primario.

Este es el principal problema que presenta este tipo de uso de los soft biometrics. Estudiando el rendimiento cuando los soft biometrics son extraídos de manera manual o automática podemos ver la diferencia en el rendimiento, esto es lo que se plantea en [30]. Cuando no se realiza ningún filtrado de los datos, la tasa de acierto es de un 98.4 %. Cuando filtramos los datos dependiendo del género y raza estimados de manera manual, los resultados mejoran ligeramente, 98.5 %. Sin embargo, cuando el género es extraído automáticamente (con un porcentaje de acierto de 84.3 %), los resultados empeoran considerablemente, 81.4 %. Resultado que empeoran más aún cuando el filtrado se basa en la raza (estimada con un acierto de 84.2 %), 73.7 %.

En [15] se estudia el rendimiento en la identificación de personas, en concreto se trata de localizar a los dos sospechosos de causar los bombardeos en el Maratón de Boston. Se utilizan dos sistemas comerciales de reconocimiento facial *NeoFace 3.1* y *PittPatt 5.2.2*. Se comparan resultados realizando una búsqueda sobre toda la base de datos y otra reduciendo las imágenes a comparar utilizando datos demográficos de los sospechosos, como la raza, el género y la edad. El número de comparaciones se reduce de un millón a alrededor de 150.000 y los resultados muestran como con los dos sistemas los índices de los rankings de identificación mejoran notablemente.

III.) Fusión con hard biometrics

Este modo de utilizar los soft biometrics, consiste en fusionar los soft biometrics, ya sean extraídos de forma manual o automática, con un sistema biométrico de primer orden, mejorando el rendimiento de este. Son múltiples los trabajos que tratan este tipo de fusión. También son múltiples las combinaciones de características y sistemas de reconocimiento biométrico que se pueden fusionar. A continuación, se repasan algunas de ellas. Como se ha comentado anteriormente el uso de los soft biometrics es de gran ayuda en escenarios complejos, por ejemplo, cuando el usuario se encuentra alejado.

En [28] se fusionan características corporales adquiridas visualmente que se usarán como las etiquetas de soft biometrics, para complementar un sistema de reconocimiento facial. Se utilizan distintos modos de fusión, todas ellas a nivel de score. Se estudian los resultados cuando el usuario se sitúa en tres distancias distintas. Se obtiene un *EER* de 15.96 % para distancias lejanas, cuando solamente se usa el sistema de reconocimiento facial (ID-SRC). Mientras que si se fusiona con las etiquetas de atributos corporales se mejora el *EER* hasta 8 % (*Sum fusion*) y 7.68 % (*Weighted fusion*). Para distancias medias y cercanas los resultados también son positivos.

También se pueden fusionar otros tipos de sistemas biométricos. Por ejemplo, en [11] se combinan distintos soft biometrics (género, altura y raza) con un sistema biométrico de reconocimiento de huella dactilar. Al tener varios soft biometrics se combinan entre ellos, teniendo distintos pesos cada una de las características, ya que por ejemplo el género aporta más información sobre la persona que la altura. De igual modo la fusión entre el sistema primario y el de soft biometrics se hace de manera ponderada. Puesto que que es más difícil falsificar la huella dactilar, el sistema biométrico primario tendrá más peso en la decisión final. Los resultados muestran que la combinación con cualquiera de los tres soft biometrics mejora la precisión del sistema de huella dactilar. El mejor resultado se obtiene cuando se usan los tres soft biometrics, se incrementa un 5 % la precisión del sistema.

En el trabajo [7] se realiza un estudio detallado para saber cuáles son los soft biometrics más discriminatorios en el reconocimiento de personas. También se analiza el rendimiento de un sistema de reconocimiento facial cuando se combina con soft biometrics extraídos manual y automáticamente. Este trabajo analiza las características de género, raza, edad, gafas, barba y bigote. Al no disponer de ninguna base de datos para la cual estas características estuvieran etiquetas, se realizó una manualmente. Esta sirvió para estudiar la discriminación de cada característica

utilizando estas como único recurso para el reconocimiento de personas. Se llega a la conclusión de que la edad es el atributo más discriminatorio, seguido de raza, género, bigote, gafas y barba. Además se demuestra que según se van usando más atributos para el reconocimiento, mejores resultados se obtienen. Se alcanza un *EER* de 11.9±2.2 cuando se usan los 6 atributos y de 11.8±2.2 cuando se incluyen todos menos la barba. La razón por la que los atributos de barba, bigote y gafas aparecen las últimas en el ranking es debido a la gran variabilidad temporal, ya que los usuarios no siempre conservan estos atributos.

El siguiente paso que se da en este trabajo es la fusión con dos sistemas de reconocimiento facial diferentes. El primero de ellos es un sistema comercial *Face++* y el segundo se trata de la red neuronal *VGGFace* [21] pre-entrenada, esta se utiliza como extractor de características. En cuanto a la extracción de los soft biometrics, se analizan dos casos por separado, cuando estos son extraídos de forma manual, haciendo uso de la base de datos *LFW* etiquetada manualmente y cuando los atributos son extraídos automáticamente mediante los sistemas *Face++* y *Microsoft Cognitive Toolkit3*. La fusión se realiza a nivel de score, promediando los soft biometrics con pesos iguales.

Para la fusión cuando los atributos son extraídos manualmente se obtienen los mejores resultados. Como se puede ver en la figura 2.2, sólo con incluir el atributo de la edad, se mejoran los dos sistemas de reconocimiento facial. Se pasa de un *EER* de 12.7 a 10.9 y de 7.8 a 7.1 con *Face++* y *VGGFace* respectivamente.

Performance of Soft Biometrics		**Face**		**Fusion**	
Set of Soft Biometrics		**Face++**	**VGG-face**	**Face++**	**VGG-face**
Age	50.6 ± 3.1	12.7 ± 1.4	7.8 ± 1.2	10.9 ± 1.4	7.1 ± 0.7
Age Ethnicity	31.1 ± 3.9			9.0 ± 1.2	5.8 ± 0.5
Age Ethnicity Gender	19.1 ± 3.3			8.4 ± 1.3	4.9 ± 0.6
Age Ethnicity Gender Moustache	14.4 ± 2.6			7.7 ± 1.5	4.8 ± 0.5
Age Ethnicity Gender Moustache Glasses	11.9 ± 2.2			7.7 ± 1.5	4.8 ± 0.7
Age Ethnicity Gender Moustache Glasses Beard	12.0 ± 2.2			8.3 ± 1.7	5.4 ± 0.9
Age Ethnicity Gender Moustache Glasses*	11.2 ± 2.1			7.6 ± 1.4	4.4 ± 0.5
Age Ethnicity Gender Moustache Glasses* Beard	11.1 ± 2.1			8.0 ± 1.7	5.2 ± 0.7

Figura 2.2: Resultados (*EER*) de utilizar soft biometrics (**obtenidos manualmente**) como *Bag of Soft Biometrics* para el reconocimiento facial *("Performance of Soft Biometrics")*. Resultados (*EER*) del reconocimiento facial utilizando los sistemas *Face++* y *VGG-Face ("Face")*. Resultados (*EER*) de la fusión de los soft biometrics con los sistemas de reconocimiento facial *("Fusion")*. *means that the sunglasses have been discarded as instance from glasses soft biometric [7].

Los resultados empeoran cuando los soft biometrics son extraídos automáticamente, como se puede er en la figura 2.3. Tanto cuando se usan como *Bag of Soft Biometrics*, que se pasa de un *EER* de 11.1 a 24.1 (cuando se usan todos los atributos); como cuando se fusionan con los sistemas de reconocimiento facial. En este caso se consigue mejorar dependiendo del número de atributos que intervengan, utilizando *Face++* son necesarios al menos tres atributos, mientras que para *VGGFace* son cuatro los atributos necesarios para mejorar. Ambos sistemas coinciden en que usando los cuatro atributos, edad, raza, género y gafas (excluyendo las de sol), se obtiene el mejor resultado, pasando de un *EER* 12.7 a 11.4 y de 7.8 a 6.6 con *Face++* y *VGG-Face* respectivamente.

Performance of Soft Biometrics		Face		Fusion	
Set of Soft Biometrics		Face++	VGG-face	Face++	VGG-face
Age	27.2 ± 1.6	12.7 ± 1.4	7.8 ± 1.2	13.7 ± 1.7	10.3 ± 1.2
Age Ethnicity	25.8 ± 2.5			12.9 ± 1.6	8.8 ± 0.7
Age Ethnicity Gender	22.2 ± 1.8			11.9 ± 1.9	8.1 ± 0.6
Age Ethnicity Gender Moustache	21.6 ± 2.0			11.7 ± 1.9	7.3 ± 0.8
Age Ethnicity Gender Moustache Glasses	22.6 ± 2.1			11.6 ± 1.9	6.8 ± 0.7
Age Ethnicity Gender Moustache Glasses Beard	23.8 ± 1.9			11.8 ± 1.8	6.9 ± 1.0
Age Ethnicity Gender Moustache Glasses*	22.7 ± 1.9			11.4 ± 1.9	6.6 ± 0.8
Age Ethnicity Gender Moustache Glasses* Beard	24.1 ± 1.7			11.5 ± 1.7	6.8 ± 1.0

Figura 2.3: Resultados (*EER*) de utilizar soft biometrics (**obtenidos automáticamente**) como *Bag of Soft Biometrics* para el reconocimiento facial *("Performance of Soft Biometrics")*. Resultados (*EER*) del reconocimiento facial utilizando los sistemas *Face++* y *VGG-Face ("Face")*. Resultados (*EER*) de la fusión de los soft biometrics con los sistemas de reconocimiento facial *("Fusion")*. *means that the sunglasses have been discarded as instance from glasses soft biometric [7].

2.3. Extracción automática de soft biometrics

Como hemos visto en el apartado anterior, la precisión con la que obtenemos los soft biometrics es muy importante, ya sea para utilizar estos como *Bag of Soft Biometrics*, o para la fusión con sistemas biométricos primarios.

Las condiciones de adquisición en las que se extraen los soft biometrics no son muy restrictivas, como sí lo son en otros sistemas. Como por ejemplo el reconocimiento de iris, para el cual son necesarios sensores muy específicos; o incluso para el reconocimiento facial, donde las oclusiones o los cambios de iluminación son cruciales para obtener buenos resultados. Esto supone una ventaja en aplicaciones de videovigilancia, o de seguimiento de personas, ya que los atributos suelen ser robustos ante cambios en las condiciones de adquisición.

El trabajo [25] aprovecha esta ventaja para reforzar la seguridad de los usuarios al usar sus dispositivos móviles. Se pretende realizar una verificación continua mediante atributos faciales, comprobando que estos son los mismos, desde el momento en el que el dispositivo se desbloquea.

Cuando el usuario desbloquea el dispositivo móvil, se adquiere una imagen frontal, para la cual se obtienen 44 atributos faciales. Durante el tiempo que el dispositivo está activo, se van adquiriendo imágenes frontales y extrayendo los soft biometrics faciales. Para la autenticación se comparan los scores obtenidos con los iniciales en el momento del desbloqueo y se procede a bloquear el dispositivo o a seguir funcionando con normalidad.

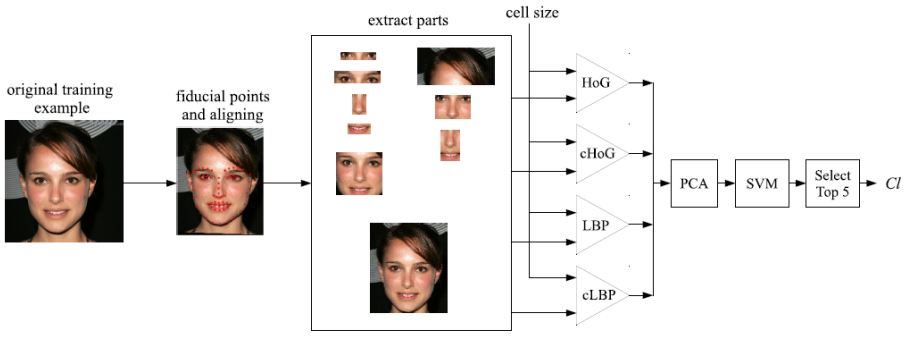

Figura 2.4: Esquema de la arquitectura levada a cabo en [25], utilizada para la extracción de atributos faciales.

Para la extracción de los soft biometrics faciales a partir de la imagen frontal obtenida, se sigue el proceso mostrado en la figura 2.4. En primer lugar, se extraen los llamados *landmarks*, que son los puntos principales de la cara, necesarios para alinear la imagen. A continuación, se divide la imagen en 9 partes, las cuales tienen asociados varios soft biometrics. Para cada parte se extraen los vectores de características a los que se les reduce la dimensionalidad mediante Análisis de Componentes Principales (PCA). Por último, se entrena un modelo SVM para cada parte y se seleccionan los 5 mejores clasificadores de cada atributo como clasificador definitivo para ese soft biometric. Para cada uno de los 44 atributos se obtienen los scores, que se utilizarán a la hora de la verificación para comprobar que el usuario que desbloqueó el móvil sigue siendo el mismo en todo momento.

Los mismos autores han ido un paso más allá y proponer mejorar el sistema anterior entrenando los modelos de extracción de soft biometrics con redes neuronales profundas adecuadas para dispositivos móviles, en vez de con clasificadores SVM. En [24] se entrena un modelo para cada parte extraída de la imagen frontal, como se observa en la figura 2.5. Se proponen dos modelos distintos, *Wide-CNNAA* y *Deep-CNNAA*. Este último tiene 5 capas convolucionales más, sin embargo las dimensiones son menores y por tanto el número de parámetros a entrenar es también menor. Los resultados mejoran el rendimineto de [25]. Además el sistema se prueba en dispositivos móviles comerciales, para evaluar el compromiso entre consumo, tiempo de ejecución y precisión en la autenticación.

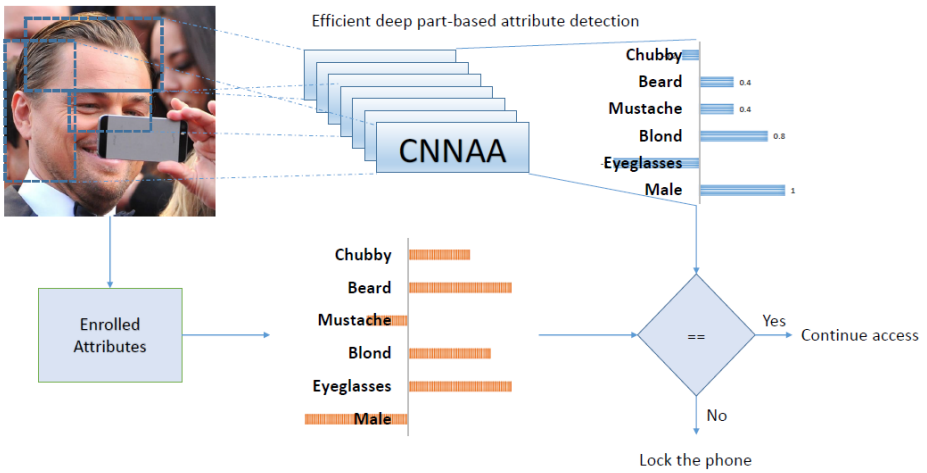

Figura 2.5: Esquema del modo de funcionamiento de la aplicación llevada a cabo en [24], para la verificación continua de personas en dispositivos móviles, mediante atributos faciales.

Muchos de los trabajos que estudian la extracción de soft biometrics, lo hacen centrándose en un atributo determinado, es decir teniendo un modelo para estudiar cada soft biometric determinado. No es el caso de [8], donde se trata de conseguir la extracción de múltiples atributos a la vez, sin la necesidad de tener un modelo entrenado para cada atributo. Para ello se hacen dos clasificaciones de los atributos. La primera de ellas atendiendo al tipo de dato, los atributos pueden ser *nominales*, cuando no hay un orden intrínseco en las distintas clases, por ejemplo la raza; u *ordinales*, cuando las distintas clases tienen un orden, por ejemplo la edad. La segunda clasificación que se hace es dependiendo del significado semántico *(semantinc meaning)*, aquí se presta atención a la región de la cara que predomina en cada atributo, se distiguen atributos *holísticos* y *locales*. Por ejemplo la edad o el género serían holísticos, ya que se describen con características de la cara completa. Mientras que atributos como la nariz puntiaguda o labios grandes se refieren a características locales de la cara y por tanto serían atributos locales.

Lo que se propone en [8] es una red de aprendizaje profunda multi tarea, del inglés *Deep Multi-task Learning (DMTL)*, para el aprendizaje de las características comunes. Esta red consiste en una modificación de la red *AlexNet*. A continuación, hay otras subredes atendiendo al tipo de atributo que se esté estudiando. Para finalizar con una última capa *FC* común a toda la red. Las funciones de coste de las subredes dependen de si el atributo es ordinal o nominal, propiedad que se asigna a la entrada de la red. El trabajo evalúa el modelo para distintas bases de datos del estado del arte, como *MORPH II*, *LFW+*, *CelebA* y *LFWA*.

2.4. Soft biometrics demográficos para reconocimiento facial

En esta sección se analizan algunos de los trabajos que utilizan atributos demográficos como ayuda en el reconocimiento facial. Son varios los artículos que demuestran que determinadas clases demográficas son más difíciles de reconocer [20], [14], [4], [6]. Según se explica en [2], el llamado *efecto raza cruzada*, en inglés *other-race effect*, demuestra que a los humanos nos cuesta más distinguir personas de otras razas que de nuestra propia raza.

Otro ejemplo de que los atributos demográficos afectan en el reconocimiento es que muchos algoritmos desarrollados en países orientale obtienen mejores resultados cuando son evaluados sobre conjuntos de personas asíaticas que occidentales. Y viceversa, los algoritmos desarrollados en occidente funcionan mejor cuando se prueban con personas occidentales [22]. Los autores de este trabajo, sugieren que este hecho es debido a los conjuntos de entrenamiento con los que han sido entrenados los sistemas.

Un trabajo muy interesante que analiza el impacto de los atributos demográficos de raza, edad y género es [14]. En esta publicación se analizan los resultados de seis sistemas de reconocimiento facial diferentes. Tres de ellos son sistemas comerciales (COTS), que proporcionan resultados similares a los del estado del arte. Estos serán cajas negras que ofrecen como salida una medida de similitud entre dos imágenes de caras. Otros dos de los sistemas son modelos de reconocimiento facial que ya han sido entrenados y por tanto no podemos modificar los conjuntos de entrenamientos. El último de los seis sistemas consiste en un modelo propio, llamado *Spectrally Sampled Structural Subspace Features (4SF)*. Lo importante de este modelo es que puede ser entrenado con distintos conjuntos de entrenamiento y analizar sus efectos en el reconocimiento facial.

Se llevan a cabo tres experimentos. El primero de ellos consiste en comparar los tres sistemas comerciales para cada grupo demográfico. El segundo experimento compara los dos sistemas no entrenables disponibles. Y por último, el tercer experimento, consiste en analizar el modelo entrenable. Para ello se entrena con cada una de las clases de cada grupo demográfico y además con un conjunto balanceado de cada grupo. Por ejemplo, para el grupo demográfico de raza, se tiene un conjunto de entrenamiento con sólo personas de raza negra, otro con personas de raza blanca, otro con personas de raza hispana y un cuarto con las tres clases balanceadas. Las evaluaciones de estos modelos se hacen sobre cada una de las clases de cada grupo demográfico. Por ejemplo, el modelo que ha sido entrenado con hispanos se evalúa sobre un conjunto de hispanos, otro de personas de raza negra y otro de raza blanca.

Algunos de los resultados de este este estudio se pueden ver en las figuras 2.6 y 2.7. Las tres primeras gráficas de estas figuras, muestran que tanto el género femenino como la raza negra son las clases que peores resultados obtienen dentro de sus respectivos grupos demográficos. Ya sea utilizando el sistema comercial, el modelo no entrenable o cuando el modelo 4SF ha sido entrenado con conjuntos balanceados. Las dos últimas gráficas de las figuras 2.6 y 2.7, así como la figura 2.8, corresponden a los resultados cuando los sistemas han sido entrenados con los conjuntos específicos de cada clase y evaluados sobre cada clase. Por ejemplo, para la raza

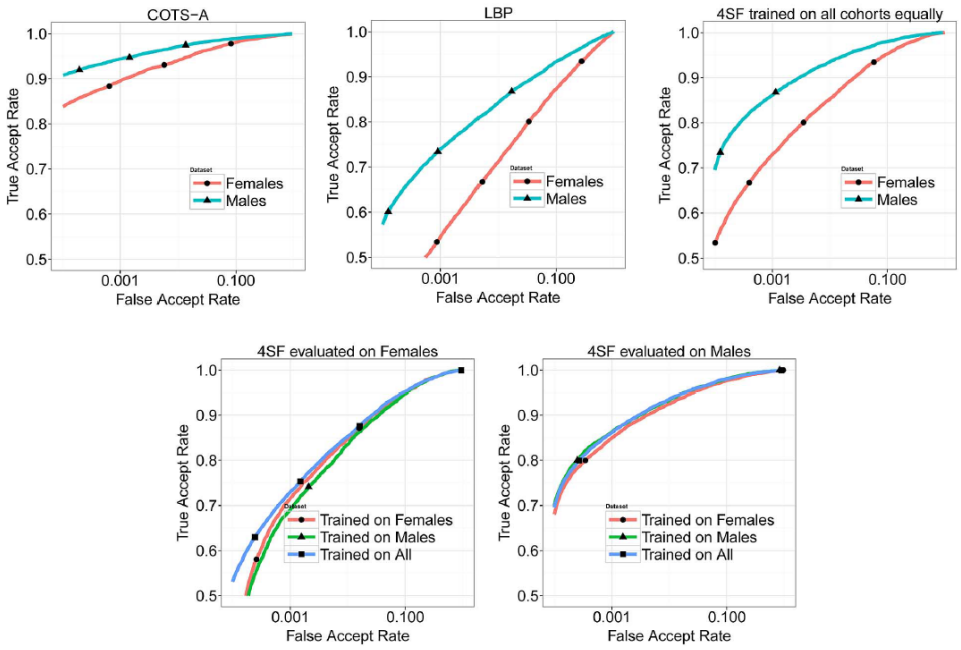

Figura 2.6: Las tres primeras gráficas comparan la precisión en el reconocimiento facial de hombres y mujeres (**género**), utilizando un sistemas comercial (*COTS*), un modelo no entrenable (*LBP*) y un modelo entrenado con un conjunto balanceado datos (*4SF*). En las últimas dos gráficas se muestran los resultados evaluando sobre sólo mujeres y sobre sólo hombres, de modelos entrenados sólo con hombres, sólo con mujeres y con un conjunto balanceado [14].

(figura 2.7), cuando el sistema es evaluado sobre un conjunto de raza blanca, se obtienen los mejores resultados cuando el sistema ha sido entrenado con un conjunto de esa misma raza, que con cualquiera de otra. Incluso cuando se entrena con el conjunto entero (cuarta gráfica figura 2.7). Lo mismo ocurre con la raza negra y con las clases 18-30 y 30-50 del grupo demográfico de edad. Para la clase de raza hispana y la de edad de 50-70, los resultados no son muy coherentes, debido al escaso número de imágenes de los conjuntos de entrenamiento.

En cuanto al género, los resultados no son tan parecidos a los otros dos atributos demográficos. Como se puede ver en la cuarta gráfica de la figura 2.6, cuando el sistema se evalúa sobre un conjunto con sólo mujeres, los mejores resultados no se obtienen cuando el sistemas ha sido entrenado con el conjunto de sólo mujeres, sino con el sistemas entrenado con todas las muestras. Según el artículo esto es debido al uso de maquillaje, haciendo que exista una alta variación en las imágenes del conjunto de mujeres, es decir, debido a la alta variabilidad intraclass.

De este artículo se sacan dos conclusiones importantes y determinantes para los siguientes capítulos de este trabajo:

- Para determinadas clases de grupos demográficos los algoritmos de reconocimiento facial funcionan peor. En concreto para mujeres, personas de raza negra y jóvenes.
- Los conjuntos de entrenamiento tienen mucha influencia en los resultados que se obtengan para cada clase. La dificultad que presentan estas clases en el reconocimiento facial, puede ser compensada cuando los modelos entrenados utilizan conjuntos de entrenamiento formados únicamente con determinadas clases. O cuando el conjunto de entrenamiento está balanceado.

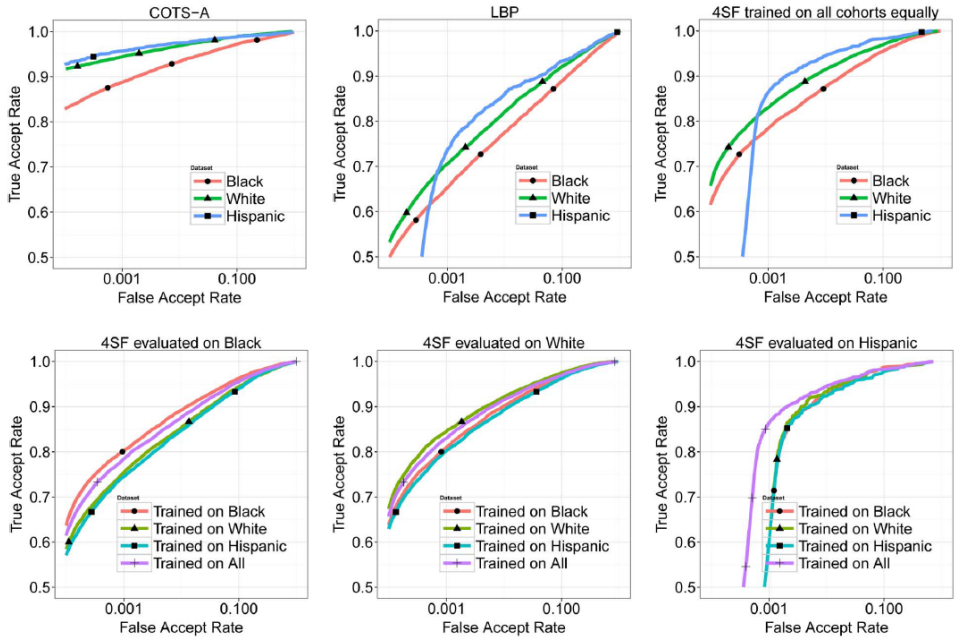

Figura 2.7: Las tres primeras gráficas comparan la precisión en el reconocimiento facial de personas de raza negra, blanca e hispana, utilizando un sistemas comercial ($COTS$), un modelo no entrenable (LBP) y un modelo entrenado con un conjunto balanceado datos ($4SF$). En las últimas tres gráficas se muestran los resultados evaluados sobre cada una de las tres clases del grupo demográfico de **raza** (negra, blanca e hispana), de modelos entrenados exclusivamente con esas tres clases y con un conjunto balanceado [14].

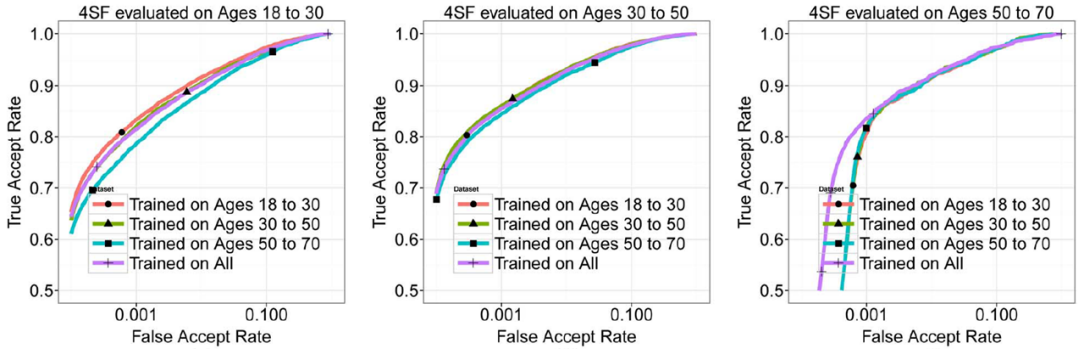

Figura 2.8: En las gráficas se muestran los resultados evaluados sobre cada una de las clases del grupo demográfico de **edad** (18-30, 30-50 y 50-70), de modelos entrenados exclusivamente con esas clases y con un conjunto balanceado [14].

3

Protocolo experimental. Bases de datos

En este capítulo se describen las bases de datos utilizadas para la realización de este trabajo, *VggFace2* y *MegaFace*. Se detalla la división que se ha llevado a cabo con cada una de ellas para crear los conjuntos de entrenamiento, validación y test. Además de describen las tareas para las que se utilizará cada una de las bases de datos. Dejando claros los conjuntos que se utilizan para cada tarea.

3.1. VggFace2

La base de datos *VggFace2* [3] contiene 3.31 millones de imágenes de 9131 identidades diferentes. Con una media de 362.6 imágenes por usuario. Las imágenes de esta base de datos tienen una gran variedad de poses, iluminación, edad y raza. Esto la convierte en una base de datos óptima para el entrenamiento de redes neuronales profundas. Además, cuenta con información adicional, como por ejemplo la posición de puntos claves de la cara, conocidos como *landmarks*, información de los *bounding boxes*, puntos necesarios para recortar la cara de cada usuario, o la información del género de cada identidad. Estas etiquetas que nos serán necesarias a la hora de entrenar los modelos específicos para hombres y para mujeres. En la figura 3.1, podemos ver algunos ejemplos de imágenes de esta base de datos después de aplicar el recorte de la cara, que es como han sido utilizadas.

Es importante tener en cuenta que esta base de datos no está balanceada ni en género ni en raza. Como se puede ver en la tabla 3.1 hay más hombres que mujeres. En cuanto a la raza, casi el 75 % de la base de datos son personas de raza caucásica. Como se verá en los capítulos posteriores este desbalanceo tendrá mucha importancia a la hora de analizar el rendimiento de los sistemas.

male	female
59.7	40.3 %

caucasian	asian	indian	african
74.2 %	6.0 %	4.0 %	15.8 %

Tabla 3.1: Distribución del número de identidades de la base de datos *VggFace2* según el género y la raza [29].

Como se verá en el siguiente capítulo, esta base de datos ha sido utilizada para dos tareas

en este trabajo:

- *Para el entrenamiento de los modelos específicos de género.* Para esta tarea necesitamos tener el máximo número de imágenes para cada usuario de cada género. Por lo que se han seleccionado 2500 hombres y 2500 mujeres, que al menos tengan 300 imágenes. De esas 300 imágenes, 240 son para el conjunto de entrenamiento y las 60 restantes para el de validación. El conjunto de test original no lo modificamos y mantenemos las 500 identidades (de ambos géneros) para nuestro conjunto de test. Estas identidades no coinciden con las de las del conjunto de entrenamiento. Únicamente se utilizan para la evaluación de los modelos, por lo que no serán vistas por la red durante el entrenamiento. La evaluación se realiza verificando si dos imágenes pertenecen o no a la misma persona. Para la validación generamos 5.000 parejas genuinas y 5.000 impostoras de cada género. Para la evaluación final de los modelos, se usarán parejas generadas pertenecientes al conjunto de test, por tanto a usuarios que no han sido vistos por los modelos durante el entrenamiento. Estas parejas de nuevo son 5.000 genuinas y 5.000 impostoras de cada género y además se añaden 5.000 impostores mixtos, sumando en total 25.000 parejas a comparar.

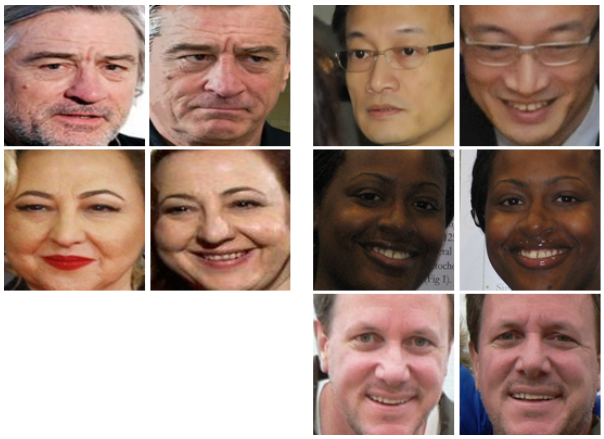

Figura 3.1: Ejemplos de imágenes de las bases de datos utilizadas. A la izquierda ejemplos de las clases hombre y mujer de la base de datos *VGGFace2* y a la derecha ejemplos de las clases hombre de raza asiática, mujer de raza negra y hombre de raza blanca de la base de datos *MegaFace*.

- *Para el entrenamiento del modelo de extracción automática del género.* En este caso, para el conjunto de entrenamiento se seleccionan 100.000 imágenes de hombres y otras 100.000 imágenes de mujeres, sin tener en cuenta la identidad de cada usuario, ya que únicamente nos interesa su género. Para el conjunto de validación se cogen 15.000 imágenes de cada género. Por último, para el conjunto de test, de nuevo se mantiene el original, de este modo los usuario con los que se evalúe el modelo no habrán sido vistos por la red durante el entrenamiento. Este contiene un total de 166.058 imágenes, de la cuales 99.684 son hombres y 66.374 mujeres. Para que el conjunto con el que se evalúa el modelo esté balanceado, usaremos únicamente 66.374 imágenes de cada género, en total 132.748 imágenes.

3.2. MegaFace

La base de datos *MegaFace* [13], contiene un total de 4.7 millones de imágenes, de 672.057 identidades diferentes, con una media de 7 imágenes por persona. En la figura 3.1 se pueden ver algunas imágenes de ejemplo. Como veremos en el capítulo siguiente, al no disponer de muchas

imágenes por usuario, esta base de datos nos servirá para entrenar modelos mediante triplet loss. Para este tipo de entrenamiento no se requiere una gran cantidad de imágenes por usuario. Esta base de datos no tiene la información de la raza ni del género. Sin embargo, mediante un proceso semi-automático se consiguió separar los 6 grupos necesarios (combinaciones de género y raza), con 4.000 identidades cada uno. Para ello se partió de la base de datos *CelebA*, de la cual se tenían los atributos de género y raza etiquetados. Se entrenó un modelo de estimación de género y otro de raza, con los que posteriormente se clasificaron las imágenes de *MegaFace*. Manualmente, se descartaron algunas imágenes mal clasificadas o que eran de mala calidad. Las imágenes que mejores resultados obtuvieron en la clasificación, fueron utilizadas para nuevamente entrenar un modelo de extracción de género y otro de raza. De nuevo se volvieron a clasificar las imágenes de *MegaFace*, para posteriormente descartar manualmente posibles errores. Hasta que cada una de las 6 clases tuviese 4.000 identidades, con al menos 3 imágenes de cada una de ellas.

De las 4000 identidades que hay en cada clase, se separan 600 usuarios para el conjunto de test. De las 3400 identidades restantes, se seleccionan 3 imágenes de cada usuario para el conjunto de validación y el resto de imágenes quedarán para el conjunto de entrenamiento. En caso de que el usuario no tenga más de 8 imágenes, no se utilizará dicho usuario, para que al menos haya 5 imágenes de cada usuario en el conjunto de entrenamiento.

Esta nueva base de datos generada a partir de *MegaFace*, se utilizará para:

- *Para el entrenamiento de los modelos específicos de raza.* Como se verá en la sección 4.1.3, estos modelos se entrenan mediante la técnica de *triplet loss*. Para ello es necesario generar trios de imágenes, estos son conocidos como *triplets*. Estos *triplets* se generaran con imágenes pertenecientes al conjunto de entrenamiento. Mientras que con las imágenes de validación, se generarán parejas genuinas e impostoras para evaluar el modelo en cada época de entrenamiento y poder utilizar el modelo óptimo. Con las imágenes del conjunto de test, se generarán parejas genuinas e impostoras y se evalúa el modelo considerado como el óptimo. En concreto, para evaluar los modelos específicos de raza, se generan 10.000 parejas genuinas y 10.000 parejas impostoras de cada raza, además de otras 10.000 parejas impostoras mixtas. En total 70.000 parejas a verificar.

- *Para el entrenamiento de los modelos específicos de género y raza (fusión).* En este caso, las clases que nos interesan son las seis originales, por lo que las utilizaremos independientemente para entrenar los seis modelos diferentes, además de un modelo mixto, que se entrena con triplets de las seis clases. Las parejas que se utilizan para evaluar estos modelos se generan a partir de las imágenes del conjunto de test. En total son 140.000 parejas a verificar, 10.000 genuinas y 10.000 impostoras de cada una de las seis clases, además de las parejas mixtas.

- *Para el entrenamiento del modelo de extracción automática de la raza.* Para este caso, se juntan las clases de ambos géneros, para tener las tres razas a clasificar. El modelo entrenado se evaluará con el conjunto de validación en cada época, para decidir la época óptima con la que finalmente se evaluará el conjunto de test, como se explicará en la sección 4.2. Este conjunto de test lo forman las 500 identidades reservadas de cada una de las seis clases. En total son 6084 imágenes de personas de raza asiática, 5706 de raza blanca y 4570 de raza negra. Como la diferencia entre cada clase no es tan notable como en el caso del género, se mantiene integro el conjunto de test, teniendo en total 16.360 imágenes.

4

Planteamiento del problema

En este capítulo, se describen los modelos desarrollados necesarios para alcanzar los objetivos de este trabajo. Así como los conjuntos de datos con los que se llevarán a cabo cada tarea.

En primer lugar, se explican las técnicas necesarias para desarrollar las tareas. Estas son la extracción de características a partir de los modelos pre-entrenados *VGGFace* y *Resnet50*. El método de fine-tuning como propuesta para el entrenamiento de los modelos específicos de género. La técnica de triplet loss, utilizada finalmente para los modelos específicos de y género y raza.

A continuación, se plantea el método llevado a cabo para los estimadores automáticos de los atributos de género y raza. Detallando la arquitectura de los modelos utilizados, los conjuntos de entrenamiento y los experimentos que se llevarán a cabo para evaluar el sistema.

Por último, se describen los sistemas propuestos para los modelos específicos de género y raza.

4.1. Modelo específico para los atributos de género y de raza

4.1.1. Extracción de características

La mayoría de sistemas de verificación tratan de comparar dos vectores de características extraídas, para decidir si dos imágenes pertenecen o no a la misma persona, como se muestra en el esquema de la figura 4.1. El preprocesamiento necesario para las imágenes que se utilizan consiste únicamente en normalizar las imágenes para que los píxeles tengan valores entre 0 y 1. El siguiente paso es recortar la cara de la persona, para que el fondo de la imagen influya lo menos posible y se tenga mayor información de la cara. En este caso la información de los puntos necesarios para proceder al recorte, llamados *bounding boxes*, se proporciona junto con las imágenes. En caso de no ser así se procedería a detectar puntos claves de la cara, estos son la situación de los ojos, boca, nariz, etc. Estos puntos son denominados *landmarks* y con ellos se puede detectar la cara de la persona y alinearla. Por último, es necesario redimensionar la imagen al tamaño de entrada a la red que vayamos a usar, en nuestro caso 224x224.

En este trabajo se han utilizado dos modelos pre-entrenados para la extracción de características, *VGGFace* [21] y *Resnet50* [9]. Estas redes están diseñadas para el reconocimiento de

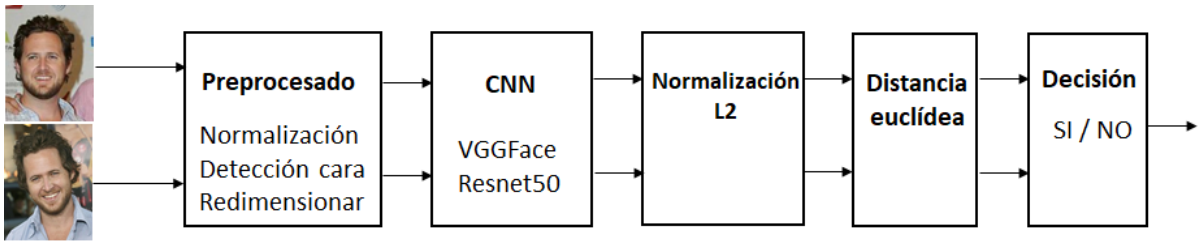

Figura 4.1: Esquema general de las diferentes etapas de un sistema de verificación.

personas, ofrecen resultados competentes con los del estado del arte y además están disponibles en [18]. Lo que las convierte en dos de las arquitecturas más utilizadas.

VGGFace es una red neuronal convolucional (*CNN*), que presenta la misma arquitectura de *VGG16* [17] (figura 4.2). Cuenta con 16 capas y un total de 145.002.878 parámetros entrenables. La red ha sido entrenada con 2.6 millones de imágenes faciales de 2622 identidades. De las cuales no se tiene información acerca de la distribución del número de personas de cada raza ni del género. Con este entrenamiento se consiguen óptimos resultados en el reconocimiento de personas, en concreto, se obtiene un 97.27 % de precisión para la base de datos LFW [10].

Resnet50 es otra red neuronal convolucional basada en una arquitectural residual, lo que la convierte en una red de menor complejidad y menor número de parámetros entrenables, 41.192.951. En la figura 4.3 se puede ver un esquema de la arquitectura de la red. Esta red ha sido entrenada con la base de datos *VGGFace2*, descrita en el apartado 3.1. Como se comenta en dicho capítulo, esta base de datos no está balanceada ni en género ni en raza. Esto es importante, porque como se verá en el capítulo siguiente los resultados son diferentes para cada clase. Una de las razones de esta diferencia es el desbalanceo del conjunto de entrenamiento con el que las redes utilizadas han sido pre-entrenadas.

Figura 4.2: Esquema de la arquitectura *VGG16* [17].

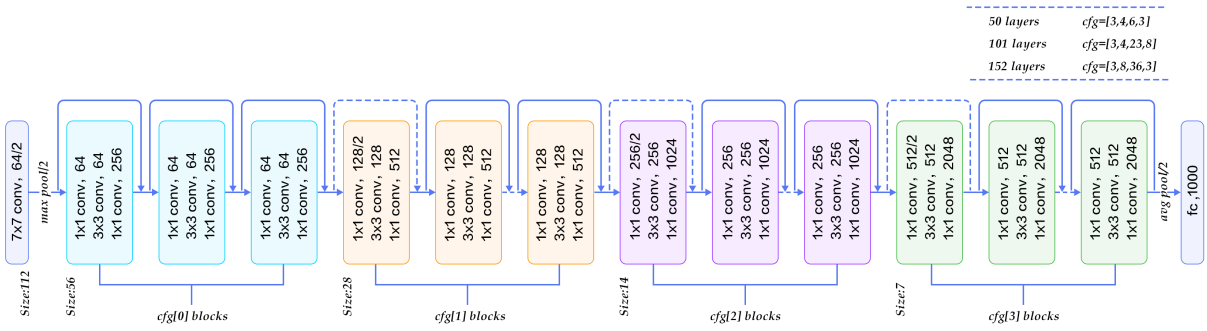

Figura 4.3: Esquema de la arquitectura *Resnet* [9].

Para obtener el vector de características a partir de la imagen preprocesada, nos hemos quedado con la salida de la capa *fc6*, para el caso de *VGGFace*, ya que se obtienen mejores resultados para clasificación que con la capa *fc7* [7]. Para el caso de *ResNet50*, nos quedamos con la salida de la capa anterior a la de clasificación. Esto hace que tengamos un vector de *L=4096* si utilizamos *VGGFace* y de *L=2048* con *ResNet50*. A continuación, normalizamos el vector mediante la normalización *L2*, para poder comparar dos vectores entre sí. Esta comparación la realizamos calculando la distancia euclídea entre ambos vectores.

4.1.2. Fine-tuning

Cuando se trabaja con modelos pre-entrenados, existen varias posibilidades de adaptar el modelo a la base de datos con la que se va a trabajar. Si tenemos una base de datos amplia, podemos re-entrenar la red desde cero. Sin embargo esta opción no es la mejor, ya que los modelos han sido pre-entrenados con bases de datos muy completas y será complicado igualar los resultados. Otra opción, es la que se conoce con el nombre de *fine-tuning*. Consiste en congelar los primeros bloques de la red, estos obtienen información común para cualquier imagen. Y son las últimas capas las que se entrenan, calculando unos nuevos pesos.

En nuestro caso, está fue la primera prueba que se hizo para entrenar los modelos específicos de género. Se pretendía entrenar mediante fine-tuning el modelo *VGGFace* y *ResNet50*, pero sólo con mujeres y de nuevo sólo con hombres. Utilizando para ello la base de datos *VGGFace2*, que como se ha explicado anteriormente, disponemos de 300 imágenes para cada identidad, teniendo 2500 mujeres para entrenar el modelo femenino y otras 2500 identidades de hombres para entrenar el modelo masculino.

En cuanto al modelo, en vez de pasar las imágenes por las capas congeladas, para reducir el tiempo de ejecución, se extrajeron y almacenaron los vectores de características para todas las imágenes. A continuación, se replicaron las últimas capas de la arquitectura que sigue el modelo *VGGFace*, es decir un capa *FC* con 4096 neuronas, seguida de otra *FC* pero en este caso con 2500 neuronas, el número de identidades a clasificar.

Para evaluar los resultados, se utilizan las parejas generadas pertenecientes al conjunto de test, como se ha explicado en el apartado 3.1. Se compararon los vectores de características de la salida de la capa anterior a la de clasificación cada para pareja y mediante funciones de la librería *sklearn*, se obtuvieron las curvas *ROC*.

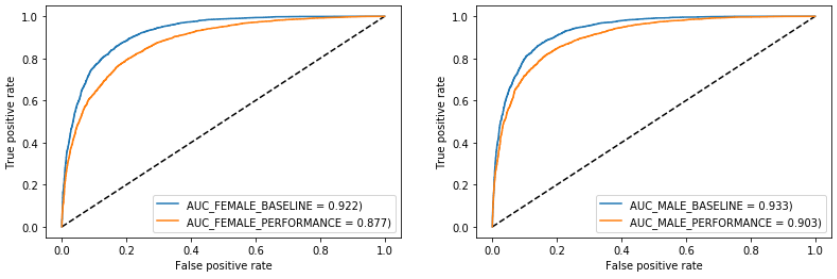

Figura 4.4: Curvas *ROC* de los modelos de género entrenados mediante *fine-tuning*. Podemos comprobar como utilizando los modelos entrenados específicamente para mujeres (izquierda) y para hombres (derecha), no mejora el rendimiento del sistema de referencia *baseline* (azul), en ninguno de los dos casos.

Los resultados obtenidos siguiendo este método no fueron los esperados, ya que como se puede ver en la figura 4.4, ni para el caso de hombres ni para el de mujeres se mejora el rendimiento del sistema de referencia, *baseline*. Lo que si que se verifica es que para el caso de hombres los

resultados son mejores que para mujeres, como se explica en varios de los artículos explicados en el estado del arte [14] [6]. La alternativa fue realizar el entrenamiento de los modelos específicos mediante triplet loss.

4.1.3. Triplet loss

La técnica de triplet loss [26] consiste en minimizar la distancia entre una imagen, a la que llamaremos *anchor* y otra de la misma identidad a la que llamaremos *positive*, frente al *anchor* y otra imagen ahora de otra identidad a la que llamaremos *negative*. Como se muestra en la figura 4.5, tras el entrenamiento con triplet loss, la distancia *anchor-positive* es menor que la distancia *anchor-negative*. Por lo tanto, si denominamos $f(x^a)$ a los embeddings de la imagen *anchor*, $f(x^p)$ a los de la imagen *positive* y $f(x^n)$ a los de la imagen *negative*. De manera que el objetivo es:

$$distanacia_{anchor-positive} << distanacia_{anchor-negative}$$

Teniendo esto en cuenta, la función de coste a minimizar durante el entrenamiento se define como:

$$L = ||f(x^a) - f(x^p)||_2^2 - ||f(x^a) - f(x^n)||_2^2 + \alpha$$

Donde α, simplemente es un margen, para que ambas distancias sean lo suficientemente distintas, suele ser de valores cercanos a 0,2.

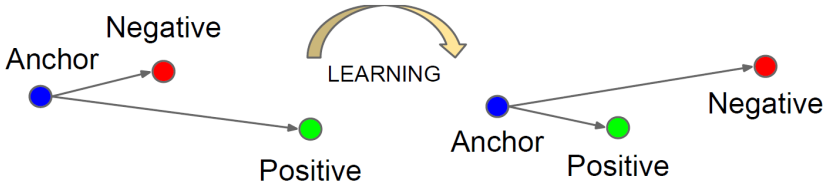

Figura 4.5: Esquema del objetivo perseguido con un entenamiento basado en triplet loss. Se pretende minimizar la distancia *anchor-positive* y maximizar la distancia *anchor-negative* [26].

Otro de los objetivos de aplicar triplet loss, es la reducción de dimensionalidad de los vectores de características. Como se ve en la figura 4.6, se pasa de tener embeddings de N=4096 o N=2048, dependiendo de si trabajmos con *VGGFace* o con *Resnet50*, respectivamente, a pasar a embeddings de M=1024. El valor de M, puede ser modificado según convenga. Simplemente es el número de neuronas de la capa de salida del modelo de triplet loss, por lo que será el tamaño de los vectores de características.

Uno de los puntos claves en el entrenamiento con triplets loss, es la elección de los triplets. Es decir, las parejas *anchor-positive* y *anchor-negative* que se van a usar para el entrenamiento. Según se explica en [26], los triplets que se usan para entrenar deberán ser lo suficientemente *difíciles* para que el modelo minimice la función de coste y aprenda durante el entrenamiento. Es decir buscamos triplets que de inicio la distancia entre *anchor-positive* sea mayor que *anchor-negative*.

Para la obtención de los triplets finales lo que hacemos es generar aleatoriamente parejas *anchor-positive* y *anchor-negative*. Posteriormente, calculamos las distancias d_{ap} y d_{an} de los embeddings extraídos usando el modelo pre-entrenado *Resnet50*. Si su diferencia supera un umbral, significa que es lo suficientemenete *difícil* y lo guardaremos para el entrenamiento. Otro factor a tener en cuenta a la hora de la selección de los triplets, es que un mismo usuario no tenga muchos triplets. Puede ocurrir que la calidad de las imágenes para un usuario no sea muy buena, por lo que la diferencia entre las distancias siempre supere el umbral. De este modo, se seleccionarían muchos triplets para ese usuario. Para solucionar este problema, limitamos el número de imágenes que un mismo usuario puede tener en los triplets a 30.

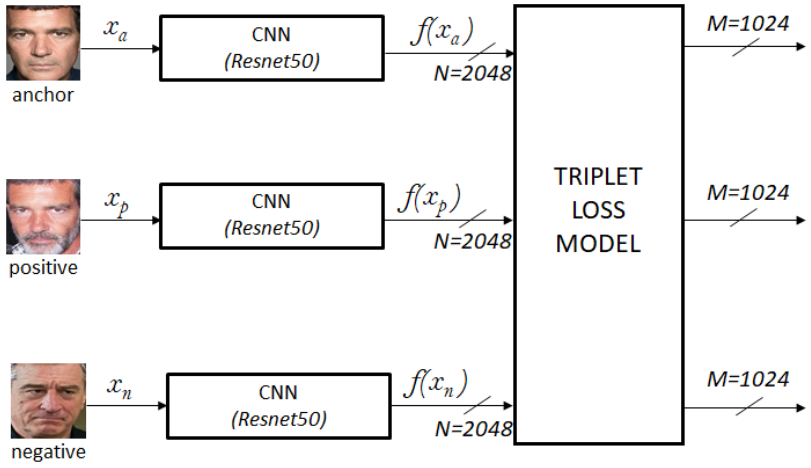

Figura 4.6: Esquema de la arquitectura del modelo de triplet loss. A partir de las imágenes *anchor*, *positive* y *negative*, se extraen los vectores de características con *Resnet50* (N=2048), para posteriormente minimizar la distancia *anchor-positive* y maximizar la distancia *anchor-negative*, además de reducir la dimensión de los embeddings (M=1024).

Como se ha explicado en el apartado 3.1, el modelo específico de género se entrena con la base de datos *VggFace2*. Se generan dos grupos de triplets uno para entrenar el modelo de hombres y otro para el de mujeres. Para cada uno de ellos, las imágenes *anchor* y *positive* serán del mismo usuario, mientras que el *negative* será de otro usuario, pero del mismo género. Para el entrenamiento del modelo de género mixto, se juntan ambos grupos de triplets. Debido a que la base de datos utilizada contiene gran cantidad de imágenes por usuario, se limitan los triplets para cada grupo, quedándonos con los 75.000 más difíciles, todos ellos superan el umbral, que en este caso se fijó en 0,2.

Para el caso de la raza, se utiliza la base de datos *MegaFace*. Del mismo modo que con el género, se generan grupos de triplets para cada clase, en este caso para asiáticos, blancos y negros. Siempre cumpliendo que las imágenes de *anchor* y *positive* sean del mismo usuario y que la de *negative* sea otro usuario, pero de la misma raza. Esta base de datos no contiene muchas imágenes por usuario, además este número es variable, para algunos usuarios hay muchas más imágenes que para otros. Para lidiar con este problema, limitamos el número de veces que una misma imagen puede ser *anchor* a 2, ya que puede ocurrir que las 30 imágenes de un usuario sean la misma. Para el caso de la raza, el umbral utilizado ha sido 0,3. De este modo el número de triplets para cada grupo ronda los 27.000. Del mismo modo que para el género, para el modelo mixto de razas, se juntal los triplets de las tres razas.

4.2. Extracción automática de los atributos de género y de raza

La información del género y la raza nos servirá en el sistema completo, para saber el modelo específico por el cual debemos pasar la imagen que estamos verificando. Para probar el sistema, podemos disponer de esta información, gracias a las etiquetas que nos proporcionan las bases de datos. Pero si queremos obtenerla de manera automática y tener así un sistema más realista, necesitamos entrenar un modelo que nos proporcione la información de manera automática y a ser posible con la mayor precisión.

Para ello se entrenan dos modelos independientes, uno para extraer la información de género y otro para la información de raza. Se ha utilizado el modelo pre-entrenado *Resnet50*, para extraer los vectores de características de todas las imágenes con las que se entrenará. Estos

embeddings, de tamaño 2048, son la entrada del modelo que entrenamos y que simplemente contiene una capa *FC* con tantas neuronas como clases (2 para el caso del género y 3 para el caso de la raza), con una activación *softmax*, para obtener como salida la probabilidad de cada clase.

Como se ha explicado en el capítulo 3, los conjuntos de entrenamiento y validación pertenecen a las mismas identidades. Ambos modelos se entrenan 70 épocas. Con el conjunto de validación se obtienen el acierto (*accuracy*) y las pérdidas (*loss*) en cada época. De este modo podemos quedarnos con la época del entrenamiento para la cual los resultados (de validación) sean mejores. Según se entena el modelo, el *accuracy* aumenta con el paso de las épocas, hasta llegar a una determinada época que comienza a descender. Con las pérdidas ocurre lo contrario, comienzan descendiendo y llega una época para la cual empiezan a aumentar. Este es el punto para el cual empieza el sobre-entrenamiento (*overfitting*). En nuestro caso damos prioridad al acierto del sistema, por los que nos quedaremos los pesos de la época para la cual los valores de *accuracy* empiecen a descender. Con esos modelos serán con los que posteriormente evaluaremos los conjuntos test. En el capítulo 5 se detallan los resultados obtenidos.

La evaluación se realiza con las imágenes de los conjuntos de test de ambas bases de datos. Para el caso del género, como se ha explicado en la sección 3.1, se seleccionan sólo 66.374 imágenes de cada género, ya que es máximo de imágenes de mujeres y de este modo realizamos la evaluación con un conjunto balanceado. Para el caso de la raza, como se detalla en la sección 3.2, se utilizan las 16.360 imágenes del conjunto de test para evaluar el modelo.

4.3. Sistema propuesto

Aprovechando la información de género y raza, se pueden hacer tres análisis interesantes. En la figura 4.7 se pueden ver los esquemas para cada atributo. A continuación, se explican cada una de las alternativas para las que en el capítulo 5 se analizarán los resultados:

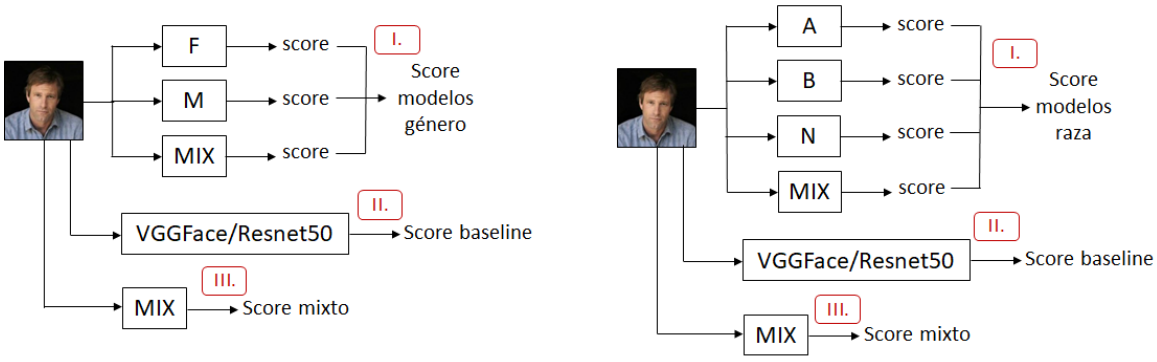

Figura 4.7: Esquema de los sistemas de género y de raza propuestos. La primera opción (*I.*) utilizando los modelos específicos entrenados. La segunda (*II.*) es la opción de referencia, utilizando directamente los embeddings extraídos de los modelos *VGGFace* y *Resnet50*. La tercera opción (*III.*) utilizando el modelo entrenado con triplets mixtos, pertenecientes a todas las clases.

- *I.* Por un lado, podemos estudiar los sistemas específicos de cada atributo. Para el caso del género, tenemos los tres modelos entrenados mediante triplet loss, uno sólo con triplets de mujeres, *modelo femenino, "F"*; otro sólo con triplets de hombres, *modelo masculino, "M"*; y un tercero entrenado con triplets de hombres y mujeres, al que llamaremos *modelo mixto, "MIX"*. Del mismo modo para la raza, tenemos el modelo para raza asiática, *"A"*;

para raza blanca, *"B"*; para raza negra, *"N"*; y un cuarto modelo entrenado con triplets de las tres razas, denominado *mixto, "MIX"*.

Para evaluar el rendimiento del sistema hemos realizado dos experimentos diferentes. El primero de ellos suponiendo que ya disponemos de las etiquetas de género y raza (suponemos que no existen errores en esas etiquetas). Y por otro lado, hemos realizado el mismo experimento pero añadiendo una etapa previa de detección automática del género y la raza, para analizar así el efecto que un sistema de detección de estos rasgos faciales puede tener en el rendimiento final del sistema de reconocimiento facial.

Para llevar a cabo la evaluación del sistema de reconocimiento facial, tendremos dos imágenes faciales que debemos comparar. Para cada una de ellas, primero obtenemos su etiqueta de género y raza (sea de forma manual o mediante el sistema automático). Si ambas son de la misma clase, es decir ambas personas son del mismo género o de la misma raza, en estos casos se empleará el modelo correspondiente en cada caso. Cuando la pareja este formada por personas de distinta clase, por ejemplo *hombre vs. mujer* o *persona de raza blanca vs. persona de raza asiática*, utilizaremos el modelo mixto en cada caso. Se calcula la distancia euclídea entre los embeddings obtenidos de cada imagen y se obtiene un score. Se agrupan los scores de todas las parejas y se obtiene el rendimiento del sistema.

Cuando la información de género o raza la obtenemos de forma automática mediante los modelos explicados en 4.2, utilizamos un umbral de confianza. El cual nos garantiza que se usará el modelo correcto. Se utilizará el modelo mixto cuando no se supere dicho umbral o cuando ambos miembros de la pareja sean de distinta clase.

- *II.* En este caso, las distancias de todas las parejas a verificar se calculan directamente con los embeddings proporcionados por los modelos *VGGFace* y *Resnet50*. Esta será la alternativa que llamaremos *baseline*, ya que es la referencia de la que se parte.

- *III.* Por último, utilizando los modelos mixtos que han sido entrenados con imágenes de los dos géneros o de las tres razas para el caso del género y la raza respectivamente. En este caso todas la parejas a verificar se pasan por los modelos mixtos (de género o de raza) y se obtienen los embeddings con los que calcular las distancias.

Para estos dos experimentos se utilizan las parejas genuinas e impostoras comentadas en el apartado 3.1. Estas son 25.000 y 70.000 para evaluar los modelos de géneros y raza respectivamente. Los resultados de estos experimentos se analizan en las secciones 5.2.3 y 5.3.3 para los casos de género y raza respectivamente.

Podemos ir un paso más allá y fusionar los atributos de género y de raza. Esta fusión se puede realizar a *nivel de features* (caso I. figura 4.8) o a *nivel de scores* (caso II. y III. figura 4.8):

- *Fusión a nivel de features.* De este modo que tenemos 7 modelos a entrenar mediante triplet loss. 6 de las posibles combinaciones de género y raza *(MA, MB, MN, HA, HB, HN)* y un séptimo que será el modelo mixto y se entrena con los triplets de todos los grupos. Del mismo modo que para el análisis del género y de la raza, para evaluar, se comprueba si ambos miembros de la pareja son de la misma clase, para estos casos se utiliza el modelo correspondiente. Por otro lado, está el modelo mixto (MIX) para los casos en los que ambos miembros de la pareja a verificar sean de distinto género y de distinta raza. Cabe destacar, que en este caso el modelo mixto será más utilizado, debido a que hay menos probabilidades de que ambos miembros de la pareja sean del mismo género y de la misma raza a la vez. Tras pasar por el modelo correspondiente, se obtienen los vectores de características y se

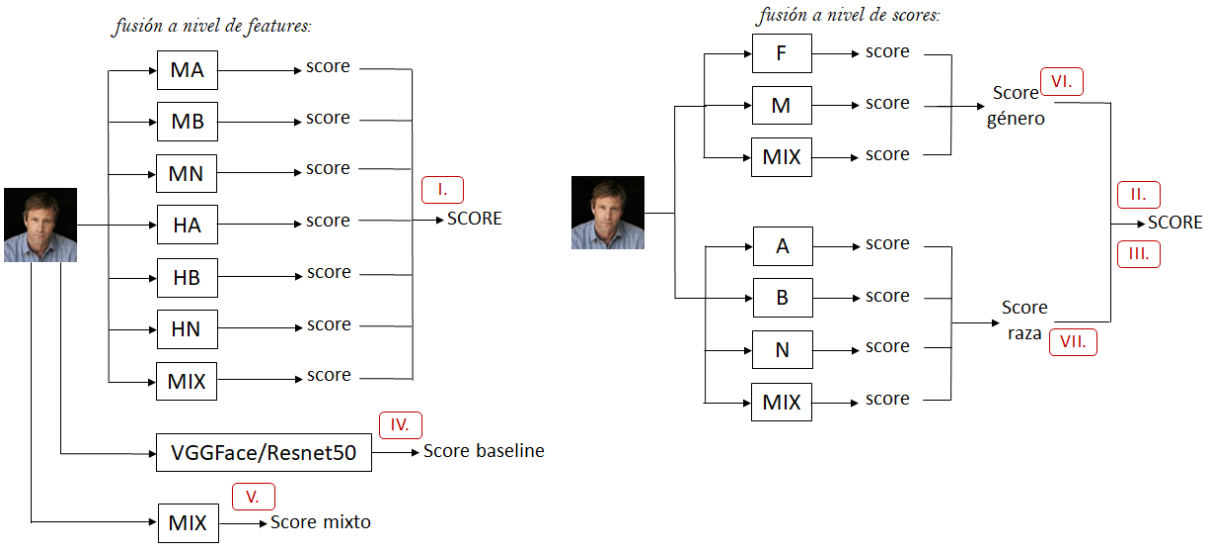

Figura 4.8: Esquema del sistema completo propuesto, fusionando información de género y de raza a nivel de features (I.) y a nivel de scores (II. y III.). Cada caja representa el modelo específico por el que se pasa la imagen, tras haber detectado la clase a la que pertenece.

calculan las distancias para obtener el rendimiento. En el esquema de la izquierda de la figura 4.8 se puede ver esta fusión.

- *Fusión a nivel de scores*. Para este caso se obtienen los embeddings utilizando los modelos específicos de género y de raza independientemente. Del mismo modo que se ha explicado al principio de esta sección y como se puede ver en el esquema de la derecha de la figura 4.8. A continuación, para fusionar ambos scores, podemos bien *promediar* (caso II.) o bien *multiplicar* (caso III.) ambos scores, para obtener los scores finales.

Las tres posibilidades de fusionar los scores, se comparan con las cuatro alternativatasiguientes:

- caso IV. Cuando todas las distancias de las parejas a verificar se obtienen directamenete de los vectores de características obtenidos a la salida de los modelos *VGGFace* y *Resnet50*. Esta es la alternativa de referencia (*baseline*).
- caso V. Cuando se utiliza el únicamente modelo mixto(entrenado con las seis clases) para todas la parejas, independientemente de su género y raza.
- caso VI. Cuando se utilizan únicamente los modelos específicos de género.
- caso VII. Cuando se utilizan únicamente los modelos específicos de raza.

La evaluación de estas siete alternativas se realiza con las 140.000 parejas de la base de datos *MegaFace*, descrita en la sección 3.2. Formadas por 10.000 parejas genuinas y 10.000 parejas impostoras para cada uno de los seis grupos, además del grupo mixto. Los resultados de este experimento se estudiarán en la sección 5.4.

Como se verá en el capítulo 5, el rendimiento de los sistemas se medirá mediante el *EER* (*Equal Error Rate*) y las *curvas ROC* (*Receiver Operating Characteristic*). El *EER* es el punto de trabajo para el cual la tasa de falsa aceptación (FAR) es igual a la tasa de falso rechazo (FRR). Por otro lado, las *curvas ROC* representan la tasa de verdaderos positivos (TAR), frente

a la tasa de falsos positivos (FAR), según se varía el umbral de discriminación. Estas se pueden resumir con el valor de *área bajo la curva* (AUC) o con valores de TAR para determinados puntos de FAR.

5

Experimentos. Resultados

En este capítulo se describen y analizan los experimentos llevados a cabo. Se han realizado los experimentos considerados para tratar de demostrar los objetivos planteados.

En primer lugar, se analizan los modelos de extracción automática de los atributos de género y raza. Primero se explica el entrenamiento llevado acabo para posteriormente exponer los resultados obtenidos.

A continuación, se analizan los resultados obtenidos con los modelos específicos de género. Lo primero que se analizan son los sistemas baseline, los cuales nos servirán como punto de partida a mejorar. También nos sirven para observar las diferencias de rendimiento que se obtienen para cada clase y con cada modelo (*VGGFace* y *Resnet50*). Posteriormente, se comparan los modelos femenino y masculino mejorados, con los modelos de referencia. Por último, se analiza el sistema completo de género, estudiando las alternativas de la obtención del género de manera manual y automática.

Seguidamente se realiza el mismo análisis para el caso de la raza. Y por último, se exponen los resultados del sistema que fusiona la información de ambos atributos. Se comparan los tres propuestas de fusión, a nivel de features y a nivel de scores mediante el promedio y mediante el producto. Así como las opciones de utilizar únicamente los sistemas de género o raza.

5.1. Extracción automática de atributos de género y raza

Con el experimento llevado a cabo en esta sección se analiza el rendimiento de los modelos entrenados para la estimación automática de los atributos de género y raza.

Como se ha explicado en la sección 4.2, con el proceso de entrenamiento llevado a cabo los modelos se entrenan las épocas suficientes, para que después con los datos de validación y teniendo en cuenta los resultados de acierto (*accuracy*) del sistema, se elija la época óptima. Tanto el modelo de raza como el de género, han sido entrenados 70 épocas. Para el caso del género, utilizaremos el modelo de la época 27, mientras que para el caso de la raza será el de la 32, ya que son en estas épocas para las que mejor *accuracy* se obtiene con los datos de validación. En la figura 5.1, podemos ver el ejemplo de las curvas de aprendizaje para el caso de la raza. A partir de la época 32, el *accuracy* empieza a descender ligeramente, lo que nos indica que

entramos en sobre-entrenamiento. Mientras que si analizamos la gráficas de las pérdidas, es casi en la época 20 cuando estas empiezan a crecer, indicándonos de nuevo donde comienza la época de sobre-entrenamiento.

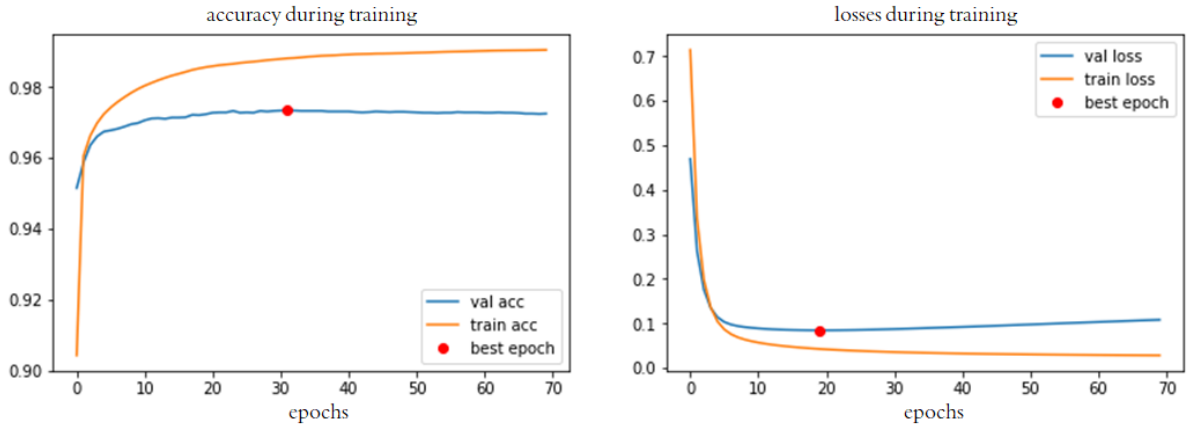

Figura 5.1: Curvas de aprendizaje para los conjuntos de validación y entrenamiento del modelo de extracción automática de raza.

Posteriormente, con los conjuntos de test, se procede a la evaluación del modelo. Estos conjuntos son los detallados en el capítulo 3, en total son 132.748 imágenes de *VggFace2* para el género y 6.084 imágenes de *MegaFace* para la raza.

Los resultados de *accuracy* y de *EER* del sistema se muestran en las tablas 5.1 y 5.2 respectivamente. Como se puede ver el modelo de extracción de género obtiene un 0.9680 de acierto. Mientras que para el caso de la raza el resultado es de 0.9777. No podemos comparar ambos resultados directamente, ya que cada uno de ellos ha sido obtenido con un número de imágenes diferentes y además estas son de bases de datos diferentes.

	género	**raza**
accuracy	0.9680	0.9777

Tabla 5.1: Resultados de *accuracy* para la extracción automática de los soft biometrics de género y raza.

	género		**raza**		
	masculino	**femenino**	**asiática**	**blanca**	**negra**
EER (%)	3.33	3.33	1.54	1.61	1.86
threshold	0.5877	0.4125	0.3044	0.4425	0.1748

Tabla 5.2: Resultados de *EER* para la extracción automática de los soft biometrics de género y raza, para cada clase (sobre los datos de test). Valores de los *thresholds* óptimos en cada caso para una óptima clasificación, obtenidos con los datos de validación.

Es importante destacar, que ambos sistemas tienen un rendimiento bastante alto, por lo que nos serán útiles para posteriormente la clasificación de los usuarios cuando la información del género y la raza se obtengan de manera automática.

Los *thresholds* que aparecen en la tabla 5.2, han sido obtenidos con imágenes del conjunto de validación y nos servirán como umbrales de confianza. Estos umbrales sirven para posteriormente cuando necesitemos obtener la información de género y de raza en las secciones 5.2.3 y 5.3.3 respectivamente, elegir el modelo a utilizar para verificar si la pareja es impostora o genuina.

En caso de que el score obtenido para una imagen no supere este umbral, se utilizará el modelo mixto, al igual que si ambas personas de la pareja a verificar, no pertenecen a la misma clase.

5.2. Género

5.2.1. Rendimiento de los sistemas baseline

Con este experimento se obtiene el rendimiento de los sistemas de referencia (*baseline*) que se pretenden mejorar con el sistema propuesto. El objetivo de este experimento es comprobar lo que muchos artículos señalan acerca de las diferencias entre hombres y mujeres en cuanto al reconocimiento facial.

De las 25.000 parejas impostoras y genuinas generadas según se detallas en la sección 3.1, se separan las 10.000 de hombres y las 10.000 de mujeres para evaluar los modelos *VGGFace* y *Resnet50*. Los resultados de pueden ver en la figura 5.2.

De este experimento hay que destacar dos cuestiones. En primer lugar, con el modelo *Resnet50* (línea continua) se obtienen mejores resultados que utilizando *VGGFace* (línea discontinua). Esto es debido a que el primero de ellos, como se explica en la sección 4.1.1, ha sido entrenado con una base de datos más amplia, con variaciones de iluminación y de pose. Además, la arquitectura de *Resent50* es más moderna y ofrece mejores resultados.

Por otro lado, es importante destacar la diferencia de resultados que se obtienen para los distintos géneros. Como se explica en varios artículos del estado del arte [14] [6], el género femenino resulta más complicado de reconocer que el masculino, esto lo podemos comprobar con este experimento. En la figura 5.2, los resultados para el género masculino están siempre por encima, tanto para el modelo *VGGFace* como para el *Resnet50*.

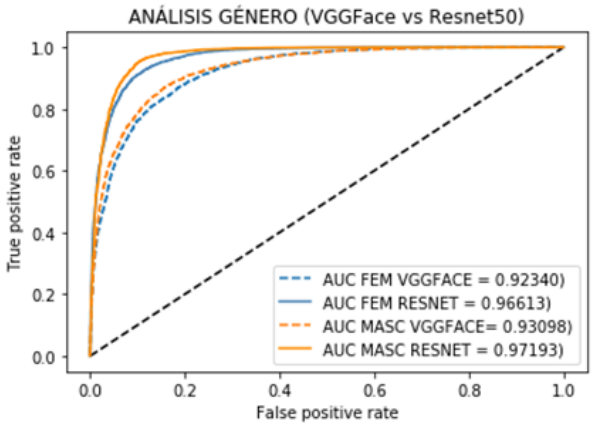

Figura 5.2: Curvas *ROC* para la verificación de mujeres (azul) y hombres (naranja), utilizando los modelos de referencia *VGGFace* (línea discontinua) y *Resnet50* (línea continua).

La tabla 5.3, muestra también estos resultados, resumiendo los valores de *EER*. Para el modelo *Resnet50* los valores de *EER* están siempre por debajo de los obtenidos con *VGGFace*. Lo mismo ocurre con el género masculino, los valores de *EER* son inferiores que para el caso de las mujeres.

Este experimento nos sirve para verificar lo que varios artículos mencionan acerca de la influencia del género a la hora del reconocimiento facial. Hemos comprobado que con un modelo pre-entrenado (*VGGFace* y *Resnet50*) con imágenes de ambos géneros, el reconocimiento para el caso de mujeres es más complejo. Hay que tener en cuenta que las bases de datos con la que estos

	género femenino		**género masculino**	
	VGGFace	**Resnet50**	**VGGFace**	**Resnet50**
EER (%)	16.1024	9.4979	14.6200	8.1400

Tabla 5.3: Resultados de *EER* para la verificación de mujeres y hombres, utilizando los modelos de referencia *VGGFace* y *Resnet50*.

modelos han sido entrenados no están balanceadas en género, hay más hombres que mujeres. Para el caso de *Resnet50* que ha sido pre-entenado con *VggFace2* sabemos que el 59.7 % son identidades masculinas [29]. Esta es una de las razones por las que los resultados para mujeres son peores. Además de la mayor variabilidad *intra-class* que existe en el conjunto de mujeres, debido al uso de maquillaje.

El siguiente paso es verificar que con un entrenamiento específico para cada género se puede mejorar el rendimiento y hacer menos notables las diferencias entre ambos géneros.

5.2.2. Rendimiento de los modelos específicos de género

En esta sección se evalúan los modelos de género entrenados mediante triplet loss. Las parejas utilizadas para este experimento son las 25.000 generadas según se explica en la sección 3.1. Para cada pareja se identifica el género (manualmente) y se utiliza el modelo correspondiente. De tal manera que el modelo femenino sólo se utilizará para parejas de mujeres y el masculino para parejas de hombres. Es decir, cada modelo es evaluado con parejas de su mismo género, del mismo modo que en [14].

El rendimiento de estos modelos se compara con los resultados de referencia del apartado anterior. Un resumen de los resultados de este experimento se puede ver en la figura 5.3 y en la tabla 5.4.

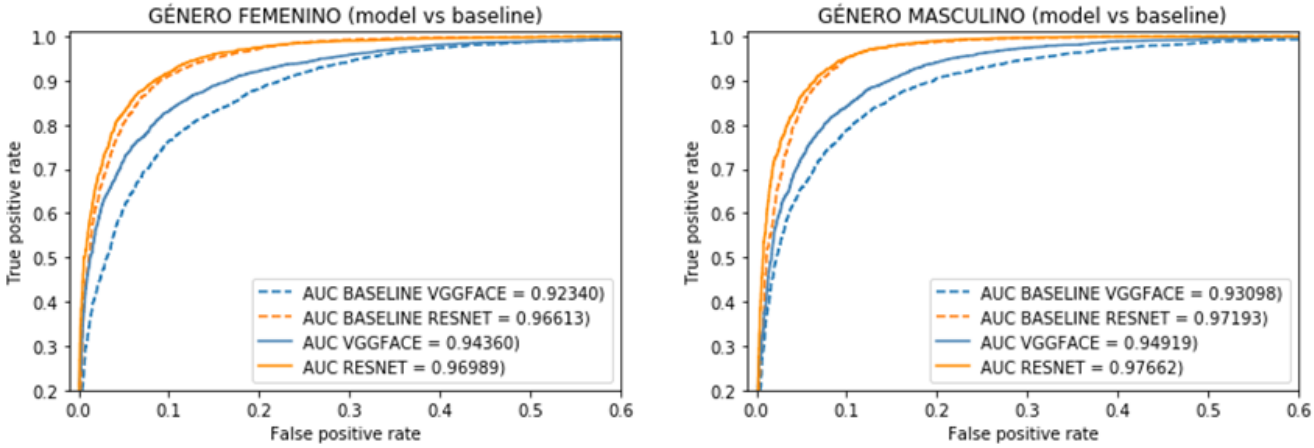

Figura 5.3: Curvas *ROC* de los modelos entrenados específicamente con mujeres (izquierda) y con hombres (derecha). Con los modelos *VGGFace* (azul) y *Resnet50* (naranja), comparados con los resultados de referencias (línea discontinua).

Para el género femenino, se consiguen mejorar los resultados de referencia, para ambos modelos (*VGGFace* y *Resnet50*). Lo podemos ver en la primera gráfica de la figura 5.3, donde además se observa que la mejora con el modelo *VGGFace* es considerablemente mejor. Los resultados de la tabla 5.4, muestran una mejora de 3 puntos de *EER* con el modelo *VGGFace*, mientras que con *Resnet50*, la mejora es de apenas 0.35. Esto es debido a que con el segundo modelo, se parte de unos resultados bastante buenos, los cuales son difícilmente superables.

	baseline		modelos	
	VGGFace	**Resnet50**	**VGGFace**	**Resnet50**
género femenino	16.1024	9.4979	13.0636	9.1440
género masculino	14.6200	8.1400	12.2000	7.6800

Tabla 5.4: Resultados de *EER* (%) para los modelos de referencia (baseline) y para los modelos entrenados específicamente de género, utilizando los modelos *VGGFace* y *Resnet50*.

Para el género masculino, ocurre algo parecido. Con ambos modelos se consigue superar los resultados de referencia y además la mejora utilizando el modelo *VGGFace* es mayor que con el modelo *Resnet50*, 2.42 frente a 0.46.

A pesar de las mejoras obtenidas para ambos géneros, sigue habiendo una clara diferencia a la hora del reconocimiento de mujeres frente a hombres. Se sigue manteniendo que el género femenino es más costoso de reconocer. Debido a la alta variabilidad *intra-class* que existe en el conjunto de mujeres. Pero sobre todo debido a los conjuntos de entrenamiento con los que los modelos de referencia han sido pre-entrenados.

5.2.3. Rendimiento del sistema propuesto con información de género

Una vez que hemos demostrado que mediante un entrenamiento específico para cada género es posible mejorar el rendimiento del sistema de referencia, podemos juntar ambos sistemas para que la evaluación se realice sobre parejas de ambos géneros. Para ello, como se explica en la sección 3.1, se evalúa utilizando las parejas genuinas e impostoras de ambos géneros y se incluyen además parejas mixtas, las cuales se pasarán por el modelo mixto, que ha sido entrenado con triplets de ambos géneros. En total son 25.000 parejas a verificar.

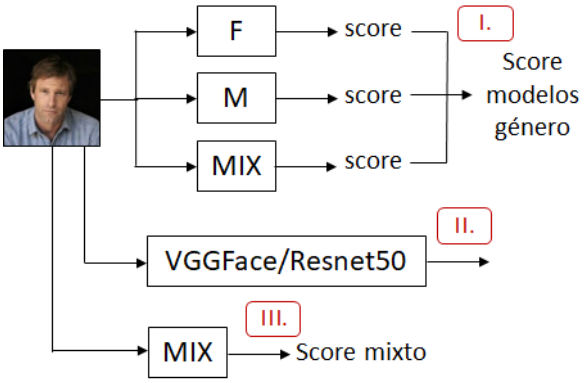

Figura 5.4: Esquema del sistema propuesto que utiliza la información de **género**.

Se siguen el esquema de la figura 5.4. La primera opción (I.) es utilizar los tres modelos entrenados, para ello previamente se detecta el género de cada miembro de la pareja, para decidir el modelo a utilizar. En caso de que cada miembro de la pareja sea de un género distinto, se emplea el modelo mixto. La segunda opción (II.) utiliza los modelos de referencia *VGGFace* y *REsenet50*. Y la última (III.) utiliza el modelo mixto para todas las parejas. El primer caso (I.) tiene la alternativa de extraer la información del género de manera automática, con el estimador propuesto en 4.2. Para este caso se emplea el *threshold* de confianza de 0.5877, para asegurarnos de que por cada modelo pasan identidades de ese género. En caso de no superar este umbral, se utiliza el modelo mixto.

En la tabla 5.5, se comparan el número de parejas que se pasan por cada modelo cuando la extracción se realiza de manera manual y automática. En el segundo caso, como es de esperar,

	# mujeres	# hombres	# mixto
manual	10000	10000	5000
automático	9586	9568	5846

Tabla 5.5: Número de identidades del conjunto de test que pasan por cada modelo cuando la extracción de la información del género se realiza de manera manual y automática.

los modelos femenino y masculino son menos utilizados, ya que nuestro extractor de género tiene un *accuracy* de 0.968, además de que no todas las imágenes superan el umbral. En concreto son 846 parejas que en vez de pasar por sus correspondientes modelos, pasan por el mixto. Esta es la razón por la que los resultados de utilizar el modelo propuesto con la extracción automática del género funcione ligeramente peor que cuando la extracción la realizamos de manera manual.

Los resultados de este experimento se muestran en la figura 5.5 y en las tablas 5.6 y 5.7. Cuando la extracción es automática se obtienen valores de *EER* de 12.14 % con *VGGFace* y 7.69 % con *Resnet50*. Mientras que si la información de género se obtiene manualmente el rendimiento mejora a 11.76 % y 7.64 % con *VGGFace* y *Resnet50* respectivamente. Ambas opciones superan al modelo de referencia (13.63 % *VGGFace* y 7.87 % *Resnet50*) y a la alternativa de utilizar únicamente el modelo mixto (12.83 % *VGGFace* y 7.88 % *Resnet50*). De nuevo ocurre que las mejoras son más notables al utilizar el modelo *VGGFace* en lugar de *Resnet50*. Ya que con el segundo de ellos, los resultados iniciales son mejores, por los que son difícilmente superables.

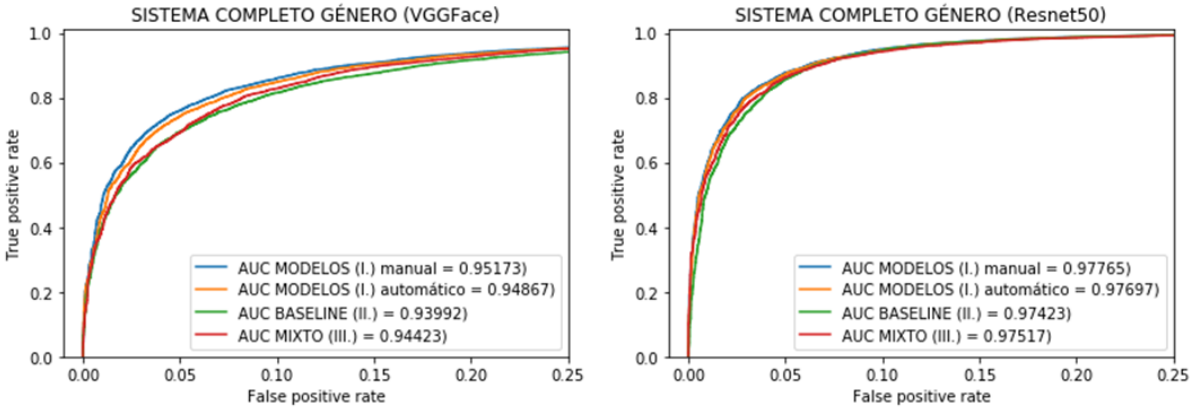

Figura 5.5: Curvas *ROC* utilizando: (I.) Sistema completo propuesto obteniendo la información del **género** manualmente (azul) o automáticamente (naranja). (II.) Modelo de referencia. (III.) Alternativa mixta, pasar todas las parejas por el modelo mixto. A la izquierda utilizando *VGGFace*, a la derecha *Resnet50*.

VGGFace	**FAR**=10^{-4}	**FAR**=10^{-3}	**FAR**=10^{-2}	**FAR**=10^{-1}	**AUC**	**EER (%)**
I. manual	0.0191	0.1735	0.4801	0.8601	0.9517	11.7600
I. auto	0.0189	0.1771	0.4389	0.8481	0.9487	12.1400
II.	0.0398	0.1139	0.4015	0.8151	0.9399	13.6300
III.	0.0292	0.1171	0.3927	0.828	0.9442	12.8300

Tabla 5.6: Resumen de resultados utilizando ***VGGFace***. (I.) Sistema completo propuesto, obteniendo la información del **género** manual o automáticamente. (II.) Modelo de referencia. (III.) Alternativa mixta, pasar todas las parejas por el modelo mixto. Se muestran: valores de *TAR*, para determinados puntos de *FAR*; el área bajo la curva *(AUC)*; y el *EER*.

Resnet50	**FAR**=10^{-4}	**FAR**=10^{-3}	**FAR**=10^{-2}	**FAR**=10^{-1}	**AUC**	**EER (%)**
I. manual	0.033	0.2111	0.6001	0.9499	0.9777	7.6436
I. auto	0.0338	0.1974	0.5891	0.9475	0.9770	7.6900
II.	6.67E-05	0.1084	0.5182	0.9462	0.9742	7.8700
III.	0.0111	0.2136	0.5661	0.9423	0.9752	7.8800

Tabla 5.7: (5.2.3) Resumen de resultados utilizando ***Resnet50***. (I.) Sistema completo propuesto, obteniendo la información del **género** manual o automáticamente. (II.) Modelo de referencia. (III.) Alternativa mixta, pasar todas las parejas por el modelo mixto. Se muestran: valores de *TAR*, para determinados puntos de *FAR*; el área bajo la curva *(AUC)*; y el *EER*.

5.3. Raza

5.3.1. Rendimiento de los sistemas baseline

Del mismo modo que hemos analizado el género podemos analizar la raza. Con este experimento observaremos las diferencias que se obtienen para cada raza, utilizando los modelos de referencia *VGGFace* y *Resnet50*. Se han utilizado las 70.000 parejas descritas en la sección 3.2. Estas se dividen según pertenezcan a cada una de las tres razas para comparar el rendimiento de cada una de ellas.

Como se puede ver en la figura 5.6, las personas de raza asiática (azul) son las que peores resultados en el reconocimineto obtienen, seguidas de las personas de raza negra (verde) y de raza blanca (naranja). En la tabla 5.8 se resumen los resultados de *EER* para este experimento. De nuevo con el modelo *Resnet50* es con el que mejores resultados se obtienen. A excepción de la raza negra, que obtiene un *EER* de 4.33 % utilizando *VGGFace*, mientras que con *Resnet50* se obtiene 4.97 %, aunque las diferencias entre ambos modelos no son tan notables, como en el caso del género.

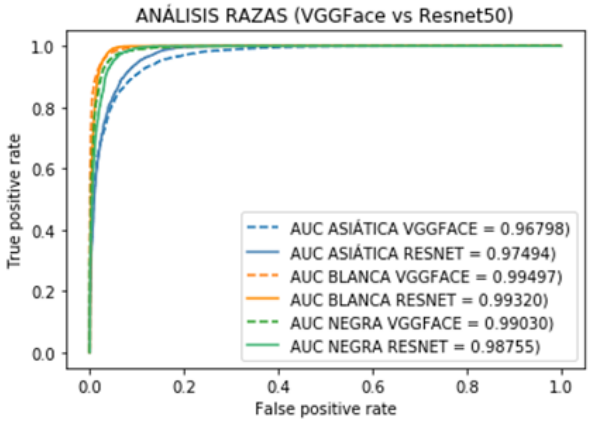

Figura 5.6: Curvas *ROC* para la verificación de personas de raza asiática (azul), raza blanca (naranja) y raza negra (verde), utilizando los modelos *VGGFace* (línea discontinua) y *Resnet50* (línea continua).

Según lo comentado en varios artículos del estado del arte [14] [6], las personas de raza negra son las que peores resultados deberían lograr. En nuestro caso no es así, esto puede ser debido a la clasificación que se ha hecho a la hora de crear la base de datos a partir de la base de datos MegaFace [13], que como se explica en la sección 3.2, se trata de un proceso semi-automático y es probable que en los grupos creados haya errores, ya que la raza en ocasiones es difícil de determinar.

Tanto este experimento como las conclusiones que se sacan de varios artículos del estado del arte coinciden en que la raza blanca es la que menos dificultad tiene en reconocimiento facial. La razón que se plantea es el hecho de que los modelos que se utilizan no están entrenados con bases de datos balanceadas. La raza predominante es la caucásica.

	raza asiática		**raza blanca**		**raza negra**	
	VGGFace	**Resnet50**	**VGGFace**	**Resnet50**	**VGGFace**	**Resnet50**
EER (%)	9.6680	8.2328	3.4215	3.4072	4.3264	4.9670

Tabla 5.8: Resultados de *EER* para la verificación de personas de las distintas razas, utilizando los modelos *VGGFace* y *Resnet50*.

5.3.2. Rendimiento de los modelos específicos de raza

Con el siguiente experimento se pretende demostrar que el entrenamiento de modelos con usuarios de una única raza mejora el rendimiento de un sistema general. Como se detalla en la sección 3.2, la evaluación de estos modelos específicos se realiza con 70.000 parejas. Estas parejas se pasarán por el modelo correspondiente a su raza. De modo que cada uno de los tres modelos es evaluado con parejas de su mismas raza. El rendimiento de cada modelo se compara con el de los sistemas de referencia utilizando las mismas parejas.

Como vemos en las gráficas de la figura 5.7 y en la tabla 5.9, se mejoran los resultados de referencia, para todos los casos. A excepción del modelo de raza blanca con *VGGFace* y el modelo de raza negra con *Resnet50*, con los que se obtiene un *EER* de 3.6 % y 5.17 % respectivamente. A diferencia de los modelos de género, aquí las mejoras no son tan notables. En el mejor de los casos, para el modelo de raza asiática con *VGGFace*, se consigue mejorar 1.18 puntos de *EER*. Una de las razones es que para los modelos de género, los resultados de referencia son peores que para el caso de la raza. Además, para el caso de la raza, no hay tanta diferencia entre los modelos de *VGGFace* y *Resnet50*.

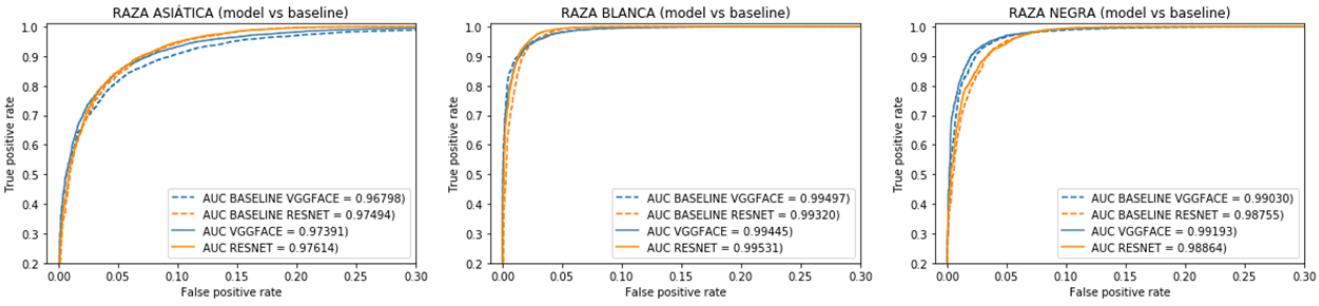

Figura 5.7: Curvas *ROC* de los modelos entrenados específicamente con personas de raza asiática (izquierda), de raza blanca (centro) y de raza negra (derecha). Con los modelos *VGGFace* (azul) y *Resnet50* (naranja). Comparando con los resultados de referencias (línea discontinua).

	baseline		**modelos**	
	VGGFace	**Resnet50**	**VGGFace**	**Resnet50**
raza asiática	9.6680	8.2328	8.4794	8.1200
raza blanca	3.4215	3.4072	3.6000	2.8223
raza negra	4.3264	4.9670	4.2000	5.1713

Tabla 5.9: Resultados de *EER* (%) para los modelos de referencia (baseline) y para los modelos entrenados específicamente de raza, utilizando los modelos *VGGFace* y *Resnet50*.

Con los nuevos modelos entrenados, se sigue manteniendo que la raza asiática es la más compleja de reconocer, seguida de la raza negra y de la blanca. Del mismo modo que sucedía con los modelos de referencia.

5.3.3. Rendimiento del sistema propuesto con información de raza

Ahora probamos los cuatro modelos entrenados, los tres de cada raza y el mixto (de raza). Para ello se utilizaran las 70.000 parejas genuinas e impostoras de la tres razas generadas según se explica en la sección 3.2. De estas parejas 20.000 son de cada una de las tres raza y las 10.000 restantes son de razas mixtas. Para estas se utilizará el modelo mixto.

El planteamiento es el mismo que para el caso del género, pero ahora siguiendo el esquema de la figura 5.8. La primera opción (I.) es utilizar la información de la raza para decidir el modelo entrenado que se va a utilizar. Teniendo la alternativa del modelo mixto para cuando cada miembro de la pareja sea de una raza diferente. La segunda opción (II.) es utilizar los modelos de referencia *VGGFace* y *Resnet50*. Y la última opción (III.) utiliza el modelo mixto para todas las parejas. Para el primer caso (I.), se propone la alternativa de obtener la información de la raza utilizando el estimador automático propuesto en 4.2. Para este caso se utilizan los tres umbrales de la tabla 5.2. De manera que si ambos miembros de la pareja son de la misma raza y además ambos superan el umbral correspondiente, se utilizará el modelo correspondiente a su raza. En caso contrario se utilizará el modelo mixto.

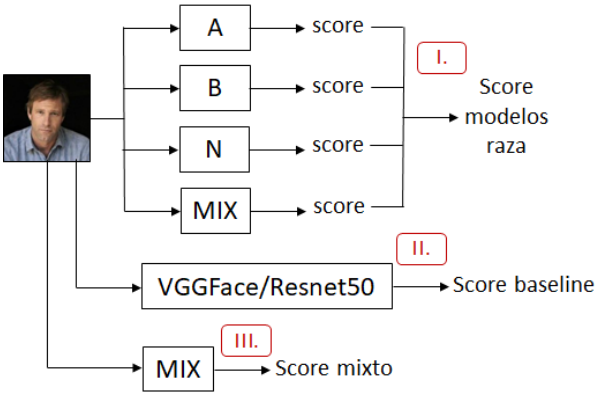

Figura 5.8: Esquema del sistema propuesto que utiliza la información de **raza**.

	# **raza asiática**	# **raza blanca**	# **raza negra**	# **mixto**
manual	21357	21241	20771	6631
automático	20752	20665	19795	8788

Tabla 5.10: Número de identidades del conjunto de test que pasan por cada modelo, cuando la extracción de la información de raza se realiza de manera manual y automática.

En la tabla 5.10 se muestra el número de imágenes que se pasan por cada modelo cuando la estimación de la raza se realiza de manera manual (sin errores) y de manera automática. Para el segundo caso se recurre más al modelo mixto, ya que no todas las imágenes superan los umbrales de confianza. Además que el estimador de raza utilizado tiene un acierto de 97.77 %. En total son 2157 las parejas que pasan por el modelo mixto en vez de por los de cada raza . Esta diferencia es la que provoca que el rendimiento para el sistema manual sea ligeramente mejor que para el automático. Como se puede ver en la tabla 5.11 los resultados de *EER* para el sistema manual y automático son de 5.60 % y 5.68 % respectivamente. Mientras que utilizando *Resnet50* (tabla 5.12) el *EER* es de 5.28 % y 5.36 % para los sistemas manual y automático respectivamente.

VGGFace	**FAR**=10^{-4}	**FAR**=10^{-3}	**FAR**=10^{-2}	**FAR**=10^{-1}	**AUC**	**EER(%)**
I. manual	0.1003	0.3583	0.7253	0.9777	0.9872	5.5967
I. auto	0.1066	0.3416	0.7158	0.9765	0.9868	5.6782
II.	0.2408	0.4465	0.7587	0.9693	0.9863	5.8616
III.	0.0510	0.3806	0.7220	0.9771	0.9869	5.6782

Tabla 5.11: Resumen de resultados utilizando ***VGGFace***. (I.) Sistema completo propuesto, obteniendo la información de la **raza** manual o automáticamente. (II.) Modelo de referencia. (III.) Alternativa mixta, pasar todas las parejas por el modelo mixto. Se muestran: valores de *TAR*, para determinados puntos de *FAR*; el área bajo la curva *(AUC)*; y el *EER*.

Los resultados de los modelos, utilizando *VGGFace* (tabla 5.11), demuestran que apenas hay diferencia entre las cuatro opciones a comparar. La mejor opción es utilizar todos los modelos, obteniendo la información de la raza de manera manual (I. (manual)). Teniendo en cuenta tanto el área bajo la curva (0.9872), como el *EER* (5.5967 %). Incluso la opción para la cual la información se obtienen de manera automática (I.(automática)) supera los resultados de referencia. Obteniendo 0.9868 de *AUC* y 5.6782 % de *EER*. Resultados prácticamente idénticos que al utilizar únicamente el modelo mixto (III.).

Resnet50	**FAR**=10^{-4}	**FAR**=10^{-3}	**FAR**=10^{-2}	**FAR**=10^{-1}	**AUC**	**EER(%)**
I. manual	0.0759	0.3305	0.6924	0.9917	0.9882	5.2764
I. auto	0.0746	0.2979	0.6701	0.9897	0.9875	5.3550
II.	0.0012	0.3173	0.6872	0.9897	0.9880	5.2333
III.	0.0298	0.2470	0.6692	0.9918	0.9881	4.9600

Tabla 5.12: Resumen de resultados utilizando ***Resnet50***. (I.) Sistema completo propuesto, obteniendo la información de la **raza** manual o automáticamente. (II.) Modelo de referencia. (III.) Alternativa mixta, pasar todas las parejas por el modelo mixto. Se muestran: valores de *TAR*, para determinados puntos de *FAR*; el área bajo la curva *(AUC)*; y el *EER*.

Para el caso de *Resnet50*, tabla 5.12, los modelos específicos no superan los resultados de referencia. Pero si lo hace la opción de utilizar únicamente el modelo mixto (III). Con el que se obtiene un *EER* de 4.96 %.

5.4. Sistema completo propuesto fusionando información de género y raza

Con los siguientes experimentos se pretende estudiar el rendimiento del sistema fusionando la información de género y raza. Para ello se estudian tres modos de fusión, a nivel de features y a nivel se scores. Esta última mediante el promedio y el producto de los scores de ambos atributos. Para la evaluación se utilizan 140.000 parejas impostoras y genuinas de cada clase (hay 6 posibles clases, combinando las de ambos atributos). En concreto son 20.000 de cada una de las 6 clases y se añaden otras 20.000 de parejas mixtas.

Se siguen el esquema de la figura 5.9. Las siete opciones que se comparan son las siguientes:

- I. Esta es la opción de fusión a nivel de features. Se utilizan los seis modelos entrenados para cada una de las clases. Estos se utilizarán cuando ambos miembros de la pareja sean de la misma clase. En caso contrario se utilizará el modelo mixto, que ha sido entrenado con triplets de las seis clases.

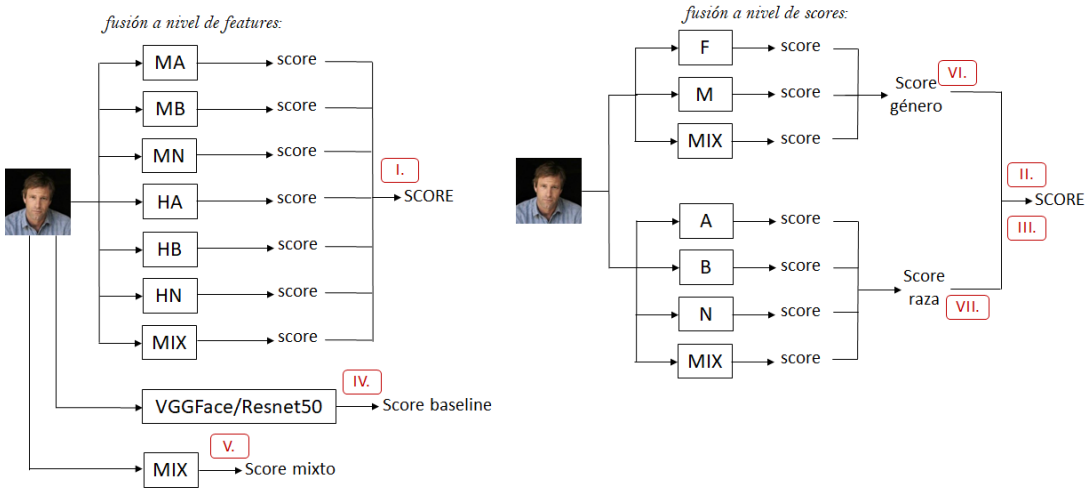

Figura 5.9: Esquema del sistema completo propuesto, fusionando información de género y de raza a nivel de features (izquierda) y a nivel de scores (derecha), como se explica en la sección 4.3. Cada caja representa el modelo específico por el que se pasa la imagen, tras haber detectado la clase a la que pertenece.

- II. En esta opción se promedian los scores de género y los de raza. Las 140.000 parejas se pasan por los modelos de género y de raza independientemente y posteriormente se promedian los scores de salida.
- III. En este caso los scores de salida de los modelos de género y de raza se multiplican para obtener los scores finales.
- IV. Para este caso se utilizan los modelos de referencia *VGGFace* y *Resnet50*.
- V. Cuando se utiliza el modelo mixto para todas las parejas, independientemente del género y la raza a la que pertenezcan.
- VI. Para este caso se utilizan únicamente los modelos de género. Previos a la fusión de scores con los modelos de raza.
- VII. En esta opción se utilizan únicamente los modelo de raza. Previos a la fusión de scores con los modelos de género.

En las tablas 5.13 y 5.14, se resumen los resultados utilizando los modelos *VGGFace* y *Resnet50* respectivamente. Se muestran los valores de *AUC* y de *EER*.

Para el caso de *VGGFace* todas las alternativas mejoran el sistema de referencia. A excepción del caso para el cual sólo se utiliza la información de género (VI.). Se obtiene un valor de *EER* demasiado alto (10.25 %), teniendo en cuenta que en la sección 5.2.3 se ha demostrado que dicho sistema consigue mejorar el rendimiento del baseline. La razón es que para esta evaluación se utilizan imágenes de *MegaFace* y los modelos de género han sido entrenados con triplets de imágenes de *VggFace2*. Sin embargo al fusionar esta información con la de raza si se mejoran los resultados, cuando se promediando o cuando se multiplican los scores. De hecho la mejor alternativa de fusionar la información de ambos atributos es promediando los scores obtenidos como salida de ambos sistemas (II.). Para esta opción se obtiene un *EER* de 6.00 %, mientras que el baseline es de 7.11 %.

Para el caso de *Resnet50*, las diferencias son también poco notables. En este caso la mejor alternativa es la fusión a nivel de features, con ella se alcanza un *EER* de 5.8330 %, mientras que

para las otras alternativas de fusión se obtiene 5.8833 % y 5.9750 % promediando y multiplicando respectivamente. De nuevo ocurre que sólo utilizando los modelos de género (7.7467 %) no se mejora el sistema de referencia (6.1850 %), pero al fusionarlos con las información de raza se consigue superar el rendimiento del baseline.

VGGFace	**(I.) features**	**(II.) promedio**	**(III.) producto**	**(IV.) baseline**	**(V.) mixto**	**(VI.) género**	**(VII.) raza**
AUC	0.9831	0.9862	0.9854	0.9814	0.9814	0.9642	0.9813
EER (%)	6.4750	6.0033	6.2150	7.1137	6.6700	10.2467	6.8563

Tabla 5.13: Resultados de *AUC* y *EER* de los distintos tipos de fusión, utilizando ***VGGFace***. Fusión a nivel de features, a nivel de scores (promediando y multiplicando), utilizando el modelo de referencia (baseline), utilizando únicamente el modelo mixto, utilizando sólo información de género y utilizando sólo información de raza

Resnet50	**(I.) features**	**(II.) promedio**	**(III.) producto**	**(IV.) baseline**	**(V.) mixto**	**(VI.) género**	**(VII.) raza**
AUC	0.9854	0.9861	0.9858	0.9846	0.9853	0.9764	0.9844
EER (%)	5.8330	5.8833	5.9750	6.1850	5.9188	7.7467	6.0500

Tabla 5.14: Resultados de *AUC* y *EER* de los distintos tipos de fusión, utilizando ***Resnet50***. Fusión a nivel de features, a nivel de scores (promediando y multiplicando), utilizando el modelo de referencia (baseline), utilizando únicamente el modelo mixto, utilizando sólo información de género y utilizando sólo información de raza

Es llamativo el rendimiento obtenido para el caso VI., cuando únicamente se utiliza la información de género. Los resultados muestran que no se supera el rendimiento del sistema de referencia. Sin embargo en el apartado 5.2.3 se ha demostrado que el sistema propuesto de género si que consigue mejorar el rendimiento baseline. Estas opciones no son comparables ya que cada una de ellas ha sido evaluada con unas parejas diferentes y además son de bases de datos diferentes. Esta puede ser la razón de la diferencia de resultados.

Otro análisis interesante que se realiza es comparar cada uno de los seis modelos específicos que utilizan la información de género y raza. Con el objetivo de examinar cual de las seis clases es la que mayor dificultad presenta. Para ello se dividen las 140.000 parejas según las seis categorías y se utiliza el modelo correspondiente para su verificación. Los resultados se comparan con los salidas de los modelos de referencia para esas mismas parejas divididas. En la figura 5.10 y en las tablas 5.15 y 5.16 se pueden ver los resultados de este experimento.

VGGFace	**MA**	**MB**	**MN**	**HA**	**HB**	**HN**
AUC	0.96047	0.9941	0.9858	0.9665	0.9930	0.9906
AUC baseline	0.9404	0.9928	0.9822	0.9644	0.9937	0.9905
EER (%)	10.6300	3.8400	5.4504	9.7700	3.7821	4.6346
EER (%) baseline	13.4693	4.1281	6.0400	10.1232	3.3900	4.6058

Tabla 5.15: Resultados de *AUC* y *EER* para los seis modelos entrenados con información de género y raza, utilizando **VGGFace**. Los resultados se comparan con el sistema de referencia (*baseline*) para cada caso.

Para *VGGFace*, las únicas clases que no mejoran el sistema de referencia son los hombres de raza blanca (3.78 %) y los hombres de raza negra (4.63 %). Sin embargo, estos valores de *EER* son muy favorables, comparados con el resto de clases. Esto puede ser debido a que la base de

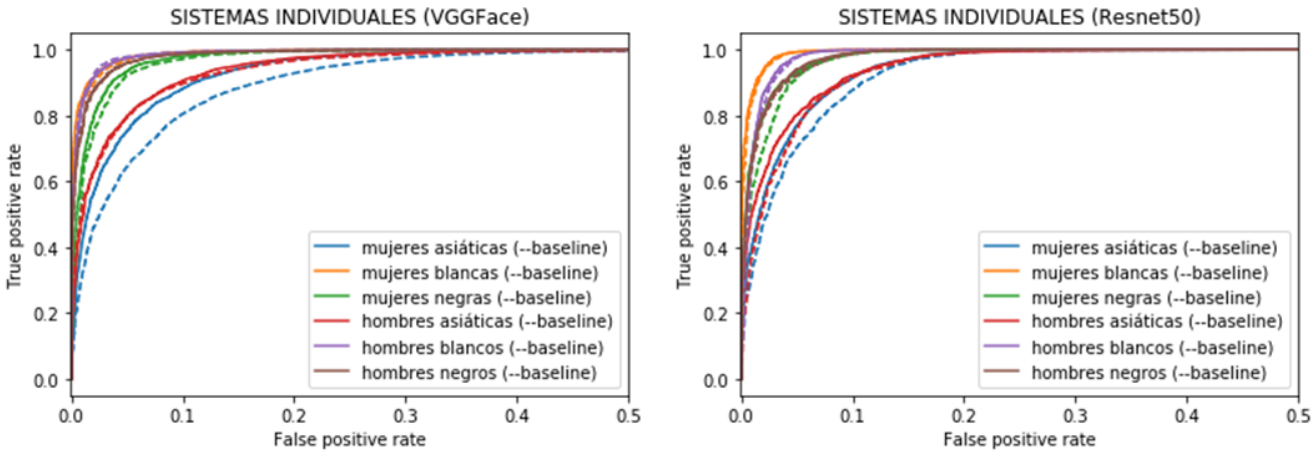

Figura 5.10: Curvas *ROC* para los seis modelos entrenados con información de género y raza, utilizando **VGGFace** (izquierda) y **Resnet50** (derecha). Cada color representa una de las categorías y en línea discontinua se representan los modelos de referencia (*baseline*).

datos con la que la red *VGGFace* ha sido pre-entrenada, contiene más hombres que mujeres. Por esta razón es más complicado superar los resultados para el caso de hombres. Del mismo modo, para las mujeres asiáticas las mejoras son las más notables, como se puede ver en la gráfica de la izquierda de la figura 5.10. La razón de nuevo puede ser por la escasez de mujeres y personas de raza asiática en el conjunto de entrenamiento utilizado para *VGGFace*.

Resnet50	**MA**	**MB**	**MN**	**HA**	**HB**	**HN**
AUC	0.9671	0.9957	0.9872	0.9708	0.9906	0.9875
AUC baseline	0.9593	0.9950	0.9830	0.9656	0.9889	0.9862
EER (%)	9.5172	2.8439	5.6997	9.1043	4.0300	5.4327
EER (%) baseline	10.6300	3.0200	5.9600	9.5000	4.2600	5.7800

Tabla 5.16: Resultados de *AUC* y *EER* para los seis modelos entrenados con información de género y raza, utilizando **Resnet50**. Los resultados se comparan con el sistema de referencia (*baseline*) para cada caso.

Para el caso de *Resnet50*, con los seis modelos se mejora el rendimiento del sistema de referencia. Las mujeres y los hombres de raza blanca son los que mejores resulados de *EER* obtienen, 2.84 % y 4.03 % respectivamente. Sin embargo, estas mejoras son muy escasas comparadas con el restos de categorías. De nuevo las mujeres asiáticas son para las que la mejora es más notable, mejoran 1.12 puntos de *EER*. Debido al desbalanceo de la base de datos *VggFace2* que es la utilizada para pre-entrenar el modelo *Resnet50*.

Con este último experimento volvemos a verificar que las personas de raza blanca son las que mejores resultados obtienen, seguidas de las personas de raza negra y asiática. Para el caso del género, son las mujeres las que mayores dificultades presentan. Estos conclusiones coinciden con las obtenidas utilizando los modelos pre-entrenados *VGGFace* y *Resnet50*, analizados en las secciones 5.2.1 y 5.3.1. Estos resultados se deben al desbalanceo en género y raza que tiene el conjunto de entrenamiento con el que los modelos han sido pre-entrenados. Recordemos que la base de datos *VggFace2*, utilizada para pre-entrenar el modelo *Resnet50*, contiene un 59.7 % de hombres y un 74.2 % de personas de raza caucásica.

6

Conclusiones y trabajo futuro

6.1. Conclusiones

Con este trabajo se ha demostrado que la incorporación de información de género y raza a un sistema de reconocimiento facial puede mejorar el rendimiento del sistema.

Para ello, en primer lugar se han demostrado las diferencias que presentan algunas clases de los grupos demográficos de género (masculino y femenino) y de raza (asiática, blanca y negra). Las mujeres y las personas de raza asiática han sido las que mayores dificultades han presentado en sistemas de reconocimiento facial. La razón principal de estas diferencias es el hecho de que los modelos empleados como extractores de características (*VGGFace* y *Resnet50*) han sido pre-entrenados con bases de datos desbalanceadas. Estas están compuestas normalmente por mayor número de hombres y de personas de raza blanca.

Para suavizar estas diferencias se han entrenado modelos para cada una de las clases de los grupos demográficos, mediante la técnica de triplet loss. Los resultados experimentales demuestran que estos modelos específicos son mejores que los de un modelo general. Se mejoran 1.87 puntos de *EER* en el mejor de los casos (utilizando *VGGFace*) para el caso del género y 0.26 puntos (utilizando *VGGFace*) para la raza.

Otra de las tareas llevadas a cabo en este trabajo ha sido la estimación automática del género y la raza, para tomar la decisión de qué modelo utilizar en cada caso. Para ello se entrenó un modelo para cada atributo, con unos resultados de *accuracy* de 96.80 % y 97.77 % para el género y la raza respectivamente.

Por último, se pretendió fusionar la información de los atributos de género y raza. Para ello se estudiaron varias alternativas de fusión, a nivel de features y a nivel de scores, esta última promediando o multiplicando los resultados obtenidos de cada atributo. Los resultados experimentales demuestran que la mejor opción es el promedio de scores para el caso de *VGGFace*, con una mejora de 1.09 puntos de *EER*. Mientras que utilizando el modelo *Resnet50*, la mejor alternativa es la fusión a nivel de features, con una mejora de 0.4 puntos de *EER*.

Al igual que muchos artículos publicados se llega a la conclusión que los conjuntos de entrenamiento tienen mucha importancia en el rendimiento del sistema. Los grupos más desbalanceados del conjunto de entrenamiento serán los que mayores dificultades tengan en el reconocimiento facial. Los hombres y las personas de raza blanca suelen obtener los resultados más favorables, coincidiendo en que son estos grupos los que normalmente predominan en las bases de datos.

6.2. Trabajo futuro

Como trabajo futuro se propone utilizar una base de datos distinta para evaluar los modelos específicos. La cual no ha sido utiliza en el entrenamiento de estos modelos. De este modo se podrán comparar mejor las mejoras de rendimiento del sistemas propuesto. Un ejemplo podría ser la base de datos *IJB* [19], que cuenta con resultados óptimos publicados.

También se podría llevar a cabo el mismo experimento para otros soft biometrics, que también son considerados discriminativos, como la edad.

Estudiar otras alternativas a la hora de tomar la decisión del modelo que se va a utilizar. Por ejemplo, al encontrarnos con una pareja hombre-mujer, utilizar el modelo específico de cada género, para posteriormente calcular las distancias, en vez de recurrir al modelo mixto, como se ha hecho en este trabajo.

Bibliografía

[1] https://www.bbc.com/mundo/noticias/2015/07/150702_tecnologia_google_perdon_confundir_afroamericanos_gorilas_lv, last access:2019-02-11.

[2] Robert K. Bothwell, John C. Brigham, and Roy S. Malpass. Cross-racial identification. *Personality and Social Psychology Bulletin*, 15(1):19–25, 1989.

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *CoRR*, abs/1710.08092, 2017.

[4] A. Dantcheva, P. Elia, and A. Ross. "What else does your biometric data reveal? a survey on soft biometrics". *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, March 2016.

[5] Antitza Dantcheva, Carmelo Velardo, Angela D'Angelo, and Jean-Luc Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, Jan 2011.

[6] Hachim El Khiyari and Harry Wechsler. Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning. *Journal of Biometrics & Biostatistics*, 7(4):1–5, 2016.

[7] Ester Gonzalez-Sosa, Julian Fierrez, Ruben Vera-Rodriguez, and Fernando Alonso-Fernandez. Facial soft biometrics for recognition in the wild: Recent works, annotation and cots evaluation. *IEEE Trans. on Information Forensics and Security*, 13(7), 2018.

[8] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2597–2609, Nov 2018.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[10] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. In *In Proc. Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France, October 2008.

[11] Anil K. Jain, Karthik Nandakumar, Xiaoguang Lu, and Unsang Park. Integrating faces, fingerprints, and soft biometric traits for user recognition. In *In Proc. ECCV Workshop BioAW*, 2004.

[12] Anil K. Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80 – 105, 2016.

[13] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *In Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4882, June 2016.

[14] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, Dec 2012.

[15] J. C. Klontz and A. K. Jain. A case study of automated face recognition: The boston marathon bombings suspects. *Computer*, 46(11):91–94, Nov 2013.

[16] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. "Describable visual attributes for face verification and image search". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, Oct 2011.

[17] S. Liu and W. Deng. Very deep convolutional neural network based image classification using small training sample size. In *In Proc. 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, Nov 2015.

[18] R. C. Malli. Github homepage. `https://github.com/rcmalli/keras-vggface`, last access:2019-02-11.

[19] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. Iarpa janus benchmark - c: Face dataset and protocol. In *In Proc. 2018 International Conference on Biometrics (ICB)*, pages 158–165, Feb 2018.

[20] G. W. Quinn P. J. Grother and P. J. Phillips. Mbe 2010: Report on the evaluation of 2d still-image face recognition algorithms.

[21] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *In Proc. British Machine Vision Conference*, 2015.

[22] P. Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J. O'Toole. An other-race effect for face recognition algorithms. *ACM Trans. Appl. Percept.*, 8(2):14:1–14:11, February 2011.

[23] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, 35(1):66–83, Jan 2018.

[24] P. Samangouei and R. Chellappa. Convolutional neural networks for attribute-based active authentication on mobile devices. In *In Proc. 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, Sep. 2016.

[25] P. Samangouei, V. M. Patel, and R. Chellappa. Attribute-based continuous user authentication on mobile devices. In *In Proc. 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, Sep. 2015.

[26] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *In Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015.

[27] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, June 2014.

[28] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. S. Nixon. Soft biometrics and their application in person recognition at a distance. *IEEE Transactions on Information Forensics and Security*, 9(3):464–475, March 2014.

[29] Mei Wang, Weihong Deng, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Racial faces in-the-wild: Reducing racial bias by deep unsupervised domain adaptation. 2018.

[30] Hao Zhang, J. Ross Beveridge, Bruce A. Draper, and P. Jonathon Phillips. On the effectiveness of soft biometrics for increasing face verification rates. *Comput. Vis. Image Underst.*, 137(C):50–62, August 2015.