Aprendizaje de representaciones faciales emocionalmente ciegas con redes neuronales profundas

Alejandro Pe¶a Almansa

Máster en Ingeniería de Telecomunicación

MÁSTERES
DE LA UAM
2019 - 2020

Escuela Politécnica Superior

UAM Universidad Autónoma de Madrid

Universidad Autónoma de Madrid

Escuela Politécnica Superior

Máster Universitario en Ingeniería de Telecomunicación

Trabajo Fin de Máster

Aprendizaje de representaciones faciales emocionalmente ciegas con redes neuronales profundas

Alejandro Peña Almansa
Tutor: Aythami Morales Moreno
Ponente: Julian Fierrez Aguilar

Septiembre 2020

Aprendizaje de representaciones faciales emocionalmente ciegas con redes neuronales profundas

Alejandro Peña Almansa
Tutor: Aythami Morales Moreno
Ponente: Julian Fierrez Aguilar

Biometrics and Data Pattern Analytics
Departamento de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Septiembre 2020

Resumen

Ante el desarrollo e implementación de los algoritmos, y su utilización en procesos de toma de decisiones automáticas en nuestra sociedad, diversos gobiernos del mundo han adoptado nuevas regulaciones con las que proteger los derechos de sus usuarios finales, tanto en cuestiones de justicia algorítmica, como en la privacidad de sus datos personales.

En este trabajo nos planteamos el potencial riesgo a la privacidad que pueden suponer tanto la monitorización, como la toma en consideración de nuestras emociones por parte de un sistema con capacidades afectivas. Por este motivo, y en el marco de las tecnologías de análisis facial, buscamos aprender nuevas representaciones faciales emocionalmente ciegas. Entendemos por representación ciega a un cierto atributo como aquella que, manteniendo la información necesaria para una tarea principal (e.g. reconocimiento facial), no permite extraer información acerca de un atributo sensible como puede ser la emoción.

Con este fin, en primer lugar llevamos a cabo una revisión del Estado del Arte en el campo del reconocimiento de emociones a partir de expresiones faciales, complementándola con una investigación de los métodos desarrollados en los últimos años para eliminar información sensible, y corregir sesgos en los algoritmos de aprendizaje automático. Tras esta revisión, proponemos dos métodos del Estado del Arte con los cuales generar las nuevas representaciones faciales emocionalmente ciegas a partir de un modelo pre-entrenado de reconocimiento facial muy popular en los últimos años.

Nuestros experimentos iniciales demuestran la presencia de información relacionada con las expresiones faciales en las representaciones extraídas por el modelo pre-entrenado, estando fuertemente embebida en estas. Mediante la adaptación y aplicación de los métodos propuestos al modelo pre-entrenado, generamos dos representaciones faciales emocionalmente ciegas, en las que la información emocional se ve significativamente reducida, a la vez que se mantiene un rendimiento considerable en verificación facial, reconocimiento de etnia y género.

Finalmente, planteamos una aplicación relacionada con la clasificación de atractivo, en la que el uso de las representaciones emocionalmente ciegas nos permite entrenar clasificadores más justos en presencia de un sesgo relacionado con la expresión facial.

Palabras Clave

Aprendizaje Automático, Aprendizaje Profundo, Reconocimiento de Patrones, Procesamiento de Imagen, Protección de Datos, Computación Afectiva

Abstract

In view of the increasing deployment of decision-making algorithms in our society, several governments around the world have adopted new regulations in this matter, which aim to protect final users, both in terms of algorithmic fairness and personal data processing.

In this work we consider the potential risk to personal privacy that may suppose systems with the capacity to both monitorice and process our emotions. For this reason, and focusing on facial analysis technologies, we seek to learn new emotional blinded face representations. A blind representation to a certain attribute is that which, while maintaining the necessary information for a main task (e.g. face recognition), doesn't allow to extract information about a sensitive attribute, such as emotion.

With this aim, we carried out a review of the State of the Art in the field of emotion recognition from facial expressions, as well as a research on the methods developed in recent years to remove sensitive information and mitigate biases in machine learning algorithms. After this investigation, we propose two different State of the Art methods to learn facial expression blinded representations, which we apply to a popular pre-trained face recognition model.

Our firsts experiments show the presence of information related to facial expression in the face representations extracted with the pre-trained model, being deeply embedded within them. By applying the two proposed methods to the pre-trained model, we generate two emotion blinded face representations, in which emotional information is significantly reduced, while maintaining competitive performance on face verification, ethnic and gender recognition.

Finally, we present an application, a case of study of attractiveness classification, in which the use of emotional blinded representations allows us to train fairer classiffiers in presence of facial expression biases.

Key words

Machine Learning, Deep Learning, Pattern Recognition, Image Processing, Data Privacy, Affective Computing

Agradecimientos

En primer lugar, me gustaría dar las gracias a todos mis compañeros del grupo de investigación BiDA Lab, que me han acompañado y acogido en este último año de estudios, y me han mostrado su apoyo constante. En este sentido, doy las gracias especialmente a Julian Fierrez y Àgata Lapedriza, por su ayuda y sus aportaciones a la realización de este trabajo, y sobre todo, a mi tutor Aythami Morales, que me ha guiado y apoyado constantemente durante su desarrollo.

También quiero dar las gracias a mis compañeros del Máster, con los que he pasado dos de los mejores años de mi carrera universitaria, así como a todos los profesores con los que compartí este periodo de mi vida.

Finalmente, a lo largo de este año el mundo entero se ha visto asolado por una grave pandemia mundial, la COVID-19. Esta enfermedad nos ha obligado a abandonar las calles, a dejar de ver a nuestra familia y amigos, a cambiar nuestro modo de vida, y sobre todo, ha provocado una gran pérdida y dolor en todos nosotros. A día de hoy, sigue existiendo una gran incertidumbre sobre el futuro que nos espera tras esta desgracia. Gran parte de la realización de este trabajo se ha desarrollado durante los meses de confinamiento, con noticias diarias sobre las pérdidas que se producían en todos los países del mundo, y sin terminar de ver la luz al final del camino. Por este motivo, me gustaría agradecer, y dedicar este trabajo, a todos los sanitarios, equipos de limpieza, cuidadores de nuestros mayores y personas dependientes, cuerpos de seguridad, investigadores, trabajadores del sector alimenticio y, en definitiva, a todas las personas que, con su esfuerzo y trabajo diario durante estos meses, han luchado por sacar al mundo adelante, y reducir al máximo las pérdidas humanas que nos dejará esta pandemia. A todos vosotros, gracias.

"Any man's death diminishes me, because I am involved in Mankind; And therefore, never send to know for whom the bell tolls; It tolls for thee."

John Donne

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Motivación

A lo largo de las últimas décadas hemos sido testigos de grandes avances en campos como el *Internet of Things* o la Inteligencia Artifical (IA), entre otros. Estos avances han sido principalmente motivados por el incremento de la capacidad computacional de la tecnología hardware, la recolección de grandes cantidades de datos de múltiples dominios de información, o el desarrollo de técnicas de reconocimiento de patrones cada vez más sofisticadas. Es gracias a estos tres motivos que progresivamente se esté generalizando el uso de algoritmos automáticos en multitud de procesos que afectan a nuestras vidas diarias, desde sistemas de recomendación [1], a procesos de selección de personal [2].

Dada la gran expansión de los algoritmos automáticos en la sociedad, son muchas las voces que han expresado su preocupación ante la posibilidad de que estos se conviertan en nuevas fuentes de discriminación (i.e. reflejando sesgos ya presentes en la sociedad [3, 4] que perjudiquen a ciertos grupos protegidos) o que supongan un riesgo para la privacidad de los usuarios. Ante esta situación, diversos gobiernos de todo el mundo han ido adoptando medidas que ponen el énfasis en prevenir la discriminación algorítmica, y en regular el procesado y almacenamiento de datos personales. En este sentido, el Reglamento General de Protección de Datos (GDPR)[1] de la Unión Europea, que entró en vigor en Mayo de 2018, es una de las regulaciones que más destacan por su impacto en los algoritmos de aprendizaje automático [5]. Esta regulación busca proteger los derechos de los ciudadanos europeos regulando cómo se procesan y almacenan sus datos personales (e.g. en los Artículos 17 y 44), introduciendo el "*right to explanation*" (e.g. en los Artículos 13 - 15) para dotar de transparencia a las decisiones tomadas por los algoritmos automáticos, o requiriendo medidas especiales para prevenir posibles efectos discriminatorios (e.g. en el Artículo 9) a causa del procesamiento de información sensible (i.e. género, etnia, creencias religiosas, etc.).

En esta línea, la comunidad científica está demostrando un gran interés en el diseño de algoritmos y herramientas que puedan prevenir la discriminación algorítmica, corri-

[1] `https://gdpr.eu/`

giendo los posibles sesgos que pueden aparecer durante el diseño de sistemas, y preservar la privacidad de los usuarios. Tópicos como "ética", "justicia" o "privacidad" son cada vez más comunes en los congresos de IA y *computer science*, siendo frecuente encontrar en estos congresos *workshops*[2] y conferencias[3] especializadas en estos temas. En este ámbito, también se puede destacar el gran trabajo que se está realizando en grandes compañías tecnológicas como Google, la cual introdujo recientemente una nueva librería en TensorFlow[4] para flexibilizar el diseño de algoritmos más "justos", o IBM, cuyo *toolkit* AI Fairness 360 [6] busca facilitar la transición del mundo de la investigación al industrial en lo referente a la justicia algorítmica.

Otro área que ha despertado gran interés en la investigación en IA de los últimos años es el conocido como *affective computing*. Este campo de la investigación tiene por objetivo el desarrollo de sistemas capaces de reconocer, interpretar, e incluso expresar emociones y sentimientos humanos [7]. Esta tecnología tiene un amplio abanico de aplicaciones, pudiendo servir para mejorar la experiencia de usuario en aplicaciones *human-centric* [8], la detección de la polaridad de los comentarios en la red (e.g. comentarios "positivos" o "negativos") [9], o para capturar la opinión pública sobre eventos sociales, movimientos políticos o campañas publicitarias [10], entre otras utilidades. Además, esta disciplina abarca diversos dominios de información, desde el mundo del *computer vision* (CV), donde típicamente se busca llevar a cabo un reconocimiento emocional a partir de expresiones faciales [11], hasta el procesamiento de lenguaje natural (NLP), en el que es destaca el conocido como *sentiment analysis* para la detección de polaridad.

Con todo esto en mente, en este Trabajo de Fin de Máster partimos de una premisa fundamental. Nuestro estado emocional es una parte vital de nuestra privacidad, y por ende, podríamos querer que no fuese analizado sin nuestro consentimiento, o que no intervenga en la toma de decisiones de los sistemas automáticos. Además, dado el aumento de los dispositivos de captura de imágenes, y el uso extendido de tecnologías de reconocimiento y análisis facial tras el éxito en estas tareas de las redes neuronales convolucionales (CNNs), ponemos el foco directamente en dichas tecnologías. Por lo tanto, el objetivo de este trabajo radica en el análisis de la información emocional (i.e. información que pueda ser explotada por un sistema para llevar a cabo tareas de reconocimiento y análisis emocional) extraída por las CNNs típicamente usadas en el ámbito del análisis facial, y el aprendizaje de nuevas representaciones faciales emocionalmente ciegas, en las que dicha información se reduzca de forma significativa sin afectar a una tarea principal.

1.2. Objetivos

Como ya se adelantó en la Sección 1.1, el objetivo principal de este Trabajo de Fin de Máster es el desarrollo de nuevas representaciones faciales, entrenadas con redes neuronales profundas (DNNs), que sean emocionalmente ciegas. De forma más desarrollada,

[2] `https://fadetrcv.github.io/`

[3] `https://www.aies-conference.com/2020/`

[4] `https://ai.googleblog.com/2020/02/setting-fairness-goals-with-tensorflow.html`

entendemos que una representación es emocionalmente ciega si la información contenida en la misma no es de utilidad a la hora de llevar a cabo tareas de reconocimiento emocional con ellas (e.g. entrenamiento de clasificadores). A su vez, nos interesa que dicha representación mantenga suficiente información para poder realizar con éxito una tarea principal distinta al reconocimiento emocional, dado que si no, esta representación no sería de especial utilidad.

Con este objetivo principal en mente, podemos definir una serie de objetivos parciales que buscamos alcanzar en este trabajo:

1. Revisión del Estado del Arte en reconocimiento de emociones, poniendo especial interés en el reconocimiento basado en expresiones faciales, cubriendo tanto técnicas como bases de datos emocionales.
2. Revisión del Estado del Arte en técnicas de eliminación de información sensible, y corrección de sesgos en redes neuronales profundas.
3. Estudio de la información emocional contenida en las representaciones faciales entrenadas con redes neuronales profundas, y típicamente utilizadas en el campo del reconocimiento y análisis facial.
4. Adaptación de métodos de eliminación de información y corrección de sesgos a la tarea de reconocimiento de emociones, generando representaciones faciales ciegas a este atributo.
5. Evaluación de las nuevas representaciones faciales, de cara a comprobar el nivel de información emocional retenido, y la utilidad de estas representaciones en otras tareas.

1.3. Organización de la memoria

Esta memoria de Trabajo de Fin de Máster está organizada en los siguientes capítulos:

- *Capítulo 1. Introducción*: En este capítulo se presentan la motivación, los objetivos y la organización general del trabajo, así como las aportaciones del mismo
- *Capítulo 2. Estado del Arte*: En este capítulo se revisan las principales tecnologías y estudios en lo referente tanto al reconocimiento de emociones a partir de expresiones faciales, como a las técnicas de eliminación de información y corrección de sesgos con redes neuronales.
- *Capítulo 3. Diseño*: En este capítulo se desarrolla formalmente el problema del aprendizaje de representaciones faciales emocionalmente ciegas, proponiendo métodos del Estado del Arte con los que aprender dichas representaciones. También se presentarán las herramientas y bases de datos usadas en este trabajo.

- *Capítulo 4. Experimentos y resultados*: A lo largo de este capítulo se presentan los distintos experimentos realizados en el marco de este trabajo, presentando y analizando los resultados de estos.
- *Capítulo 5. Conclusiones y trabajo futuro*: Este capítulo recopila las principales conclusiones de este trabajo, marcando las posibles líneas de trabajo futuro.
- *Glosario.*
- *Bibliografía.*
- *Anexos*

1.4. Contribuciones

Las principales contribuciones de este Trabajo de Fin de Máster a la comunidad investigadora son las siguientes:

- El estudio de la información emocional contenida en las representaciones faciales extraídas mediante redes neuronales profundas del Estado del Arte, comúnmente usadas en tareas de análisis facial.
- El desarrollo y evaluación de dos métodos con los que aprender representaciones faciales ciegas a las características emocionales, implementados a partir de métodos genéricos del Estado del Arte en eliminación de información sensible y corrección de sesgos.

Durante el desarrollo de este Trabajo de Fin de Máster se presentó un artículo a la conferencia internacional *International Conference on Pattern Recognition* (ICPR 2020).[5] Dicho artículo fue aceptado en Junio de 2020 en la primera ronda de revisiones, por lo que se presentará en la conferencia principal, y se publicará en las sucesivas actas. El artículo, titulado *Learning Emotional-Blinded Face Representations*, fue escrito por Alejandro Peña, Aythami Morales y Julian Fierrez de la Universidad Autónoma de Madrid, en colaboración con Àgata Lapedriza, investigadora de la Universitat Oberta de Catalunya y del *Massachusetts Institute of Technology*, siendo el estudiante y autor de este TFM el autor principal del trabajo.

[5] `https://www.micc.unifi.it/icpr2020/`

Capítulo 2

Estado del arte

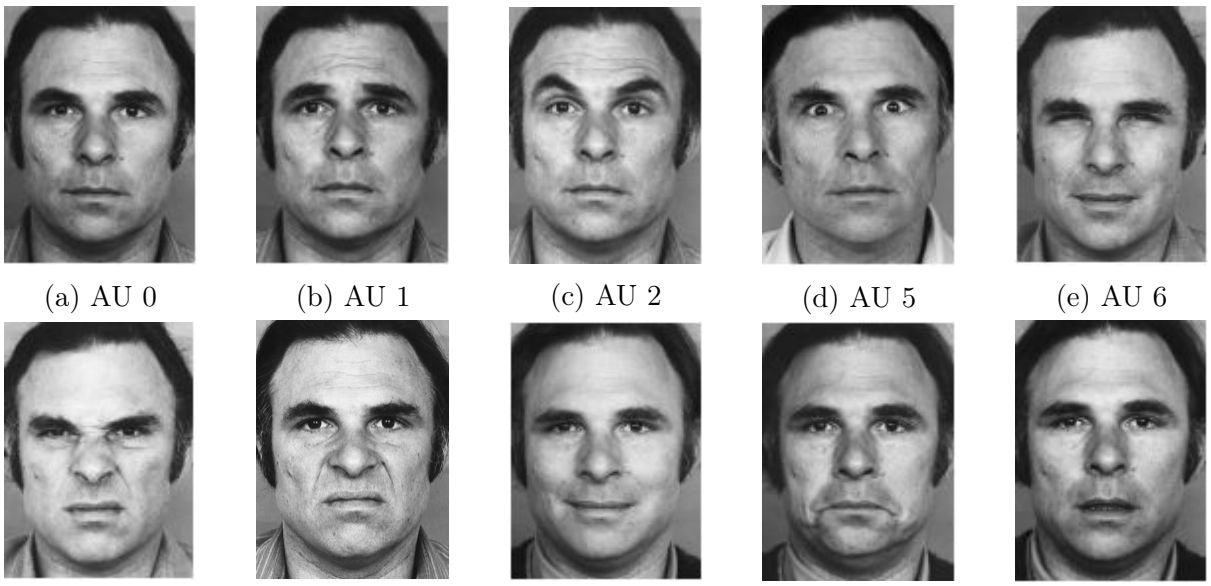
A lo largo de este capítulo realizaremos una revisión del Estado del Arte de las tecnologías que presentan relación con el contenido de este trabajo. En primer lugar, en la Sección 2.1 haremos una pequeña introducción al mundo del reconocimiento de emociones a partir de expresiones faciales. Posteriormente, en la Sección 2.2 presentaremos algunas técnicas que se han ido desarrollando en los últimos año para corregir sesgos y eliminar información sensible en los algoritmos de aprendizaje automático, con especial interés en las redes neuronales profundas.

2.1. Reconocimiento de emociones a partir de expresiones faciales

2.1.1. Sistema de Codificación Facial

La mayoría de sistemas de reconocimiento de emociones a partir de imágenes faciales se han basado a lo largo de los años en el conocido como Sistema de Codificación Facial (FACS). Originalmente propuesto en 1978 por Ekman y Friesen [12], este sistema codifica las expresiones faciales mediante un conjunto limitado de movimientos faciales, conocidos como *action units* (AUs), a los que se les asocia un cierto nivel de intensidad y una descripción. Partiendo de un conjunto de AUs que representaba movimientos principalmente localizados en la cara, con los años se fue actualizando este sistema [13] para refinar las AU existentes, e incluir movimientos más sofisticados como torsiones de la cabeza o movimientos oculares. En la Figura 2.1 se muestran varios ejemplos visuales de una selección de AUs.

A partir del sistema FACS, el cual se limita a catalogar una serie de movimientos faciales de forma objetiva, se fueron desarrollando sistemas que definen configuraciones de AU (i.e. expresiones faciales) prototípicas de cada emoción humana. Este es el caso del Sistema de Codificación Facial Emocional (EMFACS) [14], que incluye este nivel de subjetividad para asociar AUs y emociones. En la Tabla 2.1 se recogen las asociaciones más



(a) AU 0 (b) AU 1 (c) AU 2 (d) AU 5 (e) AU 6

(f) AU 9 (g) AU 10 (h) AU 12 (i) AU 15 (j) AU 25

Figura 2.1: Ejemplos visuales[1] de una selección de AUs definidas por el FACS [12, 13]. Los movimientos faciales representados por cada AU son: (a) neutral, (b) levantamiento interior de ceja, (c) levantamiento exterior de ceja, (d) levantamiento del párpado superior, (e) levantamiento de mejillas, (f) arrugar la nariz, (g) levantamiento del labio superior, (h) levantamiento labial esquinal, (i) depresión labial esquinal y (j) deslizamiento labial.

comunes para las 6 emociones principales definidas por Ekman y Friesen [15], existiendo trabajos que presentan ciertas modificaciones en las mismas.

Dada esta asociación, muchos sistemas han usado la detección de AUs como método para llevar a cabo el reconocimiento de este conjunto de emociones. De esta forma, si en una imagen facial se detecta el patrón de AUs correspondiente a una emoción determinada, puede clasificarse como tal. Este tipo de clasificación asume que cada una de las emociones básicas se puede expresar de forma universal con una expresión facial determinada. Sin embargo, algunos estudios psicológicos y culturales más recientes aseguran que esta relación no es universal [16, 17], sino que es específica de cada cultura, e incluso puede no ser de utilidad en pacientes con desórdenes neurodegenerativos como puede ser la enfermedad de Parkinson [18]. A pesar de estas limitaciones, la detección de emociones a partir de expresiones faciales ha sido uno de los modelos más usados en reconocimiento de emociones.

2.1.2. Técnicas de reconocimiento de emociones a partir de expresiones faciales

Al igual que en otras tareas típicas dentro del mundo del CV, el reconocimiento de emociones a partir de expresiones faciales ha ido evolucionando en la última década des-

[1] `http://cbcsl.ece.ohio-state.edu/enc-2020/index.html`

Emoción	AU	Emoción	AU
Felicidad	$6 + 12$	Miedo	$1 + 2 + 4 + 5 + 7 + 20 + 26$
Tristeza	$1 + 4 + 15$	Sorpresa	$1 + 2 + 5 + 26$
Enfado	$9 + 15 + 16$	Disgusto	$4 + 5 + 7 + 23$

Tabla 2.1: Asociación de las emociones principales con expresiones faciales, en términos de AU según el EMFACS [14].

de sistemas basados en características *handcrafted* (i.e. características diseñadas por el ingeniero) a sistemas que utilizan DNNs tanto para la extracción de características, como para la clasificación a partir de las mismas. Históricamente, una de las mayores limitaciones que ha experimentado el campo es la escasez de datos etiquetados con los que llevar a cabo el diseño de estos sistemas [11], lo que limitó su desarrollo a entornos de laboratorio controlados. Sin embargo, a partir de 2013 las competiciones de reconocimiento de emociones (e.g. FER-2013 [19] o EmotiW [20, 21]) comenzaron a recolectar grandes bases de datos, lo que permitió a estos sistemas enfrentarse a desafíos más realistas, y comenzar su transición a los métodos *deep learning*, obteniendo resultados nunca vistos en este campo. A día de hoy, existen múltiples bases de datos etiquetadas con información referente a la expresión facial y la emoción, disponibles para el entrenamiento de sistemas de análisis emocional. La Tabla 2.2 incluye una selección de las principales bases de datos de los últimos años, con información relevante de cada una de ellas.

2.1.2.1. Reconocimiento de expresiones faciales mediante métodos tradicionales

Entendemos por "método tradicional" de reconocimiento de expresiones faciales como un sistema que no se basa en técnicas de *deep learning* para llevar a cabo dicha tarea, sino que se basan en el uso de características *handcrafted*. Estos métodos, por lo tanto, fueron típicamente usados antes del éxito de las DNNs de la última década, e incluían una gran variedad de aproximaciones.

Son varios los trabajos que se basan en la extracción de patrones locales binarios (LBP) como base para llevar a cabo una clasificación de expresiones faciales independiente del individuo [22], dada su simplicidad y su éxito en tareas de reconocimiento facial. En [22] se propone la extracción de las características LBP más discriminativas mediante el uso de AdaBoost, y su posterior clasificación usando máquinas de vectores soporte (SVM). Los autores de [23] proponen una extensión de LBP para llevar a cabo un reconocimiento dinámico de texturas con aplicación al reconocimiento de expresiones en secuencias, calculando y concatenando características LBP en 3 planos ortogonales para incluir información de apariencia y movimiento.

En [31] se comparan diversas técnicas de clasificación, incluyendo algoritmos como AdaBoost, SVM o análisis discriminante lineal, así como métodos de selección de características para reconocimiento de expresiones faciales. Los resultados de dicho trabajo mostraban el mejor rendimiento al seleccionar un conjunto de filtros de texturas de Gabor

Nombre	Dataset	Identidades	Fuente	Etiquetas
EmotioNet [24]	1M imágenes	N/A	Internet	23 AU
CFEE [25]	6K imágenes	230	Lab	26 emociones
CK+ [26]	593 secuencias	123	Lab	30 AU y 7 emociones
FER-2013 [19]	36K imágenes	N/A	Internet	7 emociones
DISFA [27]	130K imágenes	27	Lab	12 AU
SFEW 2.0 [21]	1,635 imágenes	N/A	Películas	7 emociones
AFEW 5.0 [21]	1,645 vídeos	N/A	Películas	7 emociones
AFEW 7.0 [28]	1,809 vídeos	N/A	Películas	7 emociones
AffectNet [29]	450K imágenes	N/A	Internet	8 emociones con valencia
MultiPie [30]	750K imágenes	337	Lab	6 emociones

Tabla 2.2: Selección de bases de datos emocionales de la última década, utilizadas para el reconocimiento de emociones y expresiones faciales.

mediante el algoritmo de AdaBoost para extraer características, seguido de una SVM en clasificación. Otra aproximación es la propuesta en [32], donde se reconocen AUs y sus combinaciones mediante razonamiento basado en reglas a partir de la detección de contornos faciales. También se ha llevado a cabo esta tarea mediante factorización matricial no negativa [33] o *sparse learning* [34].

Como apunte final de esta sección, algunos trabajos buscaban explotar el comportamiento temporal de las expresiones faciales, trabajando sobre secuencias de imágenes directamente. En este ámbito, las principales aproximaciones se basaban en el uso de modelos ocultos de Márkov (HMM), tanto para la clasificación de secuencias [35, 36] como para su segmentación [35], así como redes bayesianas dinámicas (DBN) [37].

2.1.2.2. Reconocimiento de expresiones faciales basados en deep learning

Los métodos de reconocimiento de expresiones basados en técnicas de *deep learning* empezaron a ganar relevancia a comienzos de la última década, demostrando rápidamente un rendimiento muy superior a sus predecesores. Aunque el uso de redes neuronales (NNs) ya había sido propuesto años atrás en este ámbito, no es hasta 2013 cuando dos propuestas basadas en CNNs fueron las ganadoras en dos de las principales competiciones de reconocimiento de expresiones faciales. En la primera, FER-2013 [19], el trabajo ganador [38] proponía el uso de una CNN reemplazando la capa *softmax* de clasificación por una SVM lineal, de forma que la función de pérdidas que se minimizaba era una *Hinge loss* en vez la clásica *cross-entropy loss*. Por su parte, el trabajo ganador de EmotiW 2013 [20] proponía un sistema multimodal [39] que agregaba una CNN para el análisis de expresiones faciales en el vídeo, una *deep belief network* (DBN) para el análisis de audio y un *deep autoencoder* para extraer información espacio temporal.

De forma genérica, los sistemas de reconocimiento de expresiones faciales basados en *deep learning* suelen incluir una etapa de preprocesamiento de las imágenes al trabajar con entornos no controlados. En esta etapa, al menos es indispensable la presencia de un

sistema de detección facial, como puede ser el detector Viola-Jones [40]. Con el objetivo de mejorar el rendimiento, también es común durante este procesado llevar a cabo una alineación facial mediante la extracción de puntos de referencia faciales. Existen múltiples aproximaciones para llevar a cabo dicha alineación, como pueden ser los detectores basados en partes [41], la aplicación de múltiples regresores en cascada [42], o el uso de CNNs en cascada [43]. En esta última línea, también se han utilizado algoritmos que combinan tanto la detección como la alineación facial (e.g. el detector MTCNN [44]).

Debido a la escasez de datos en este dominio, el entrenamiento directo de CNNs usando bases de datos de expresiones faciales corre el riesgo de acabar en *overfitting* [11]. Con el objetivo de evitar este efecto, muchos trabajos utilizan datasets adicionales de mayor tamaño para llevar a cabo un *pre-training* de la red, o recurren a un *fine-tuning* de modelos ya entrenados. En [45] se expone una estrategia multietapa, partiendo de una red pre-entrenada, a la que se aplica un primer *fine-tuning* genérico con el dataset FER-2013 [19], para posteriormente refinarlo usando el dataset objetivo. Los resultados de [46] sugieren que realizar un *fine-tuning* de una red pre-entrenada para reconocimiento facial con un dataset de expresiones faciales incrementa la precisión en el reconocimiento de estas, solventando el problema de la escasez de datos. Sin embargo, en este proceso las características extraídas por la red pueden contener información del dominio original que perjudique el reconocimiento de emociones. Con esto en mente, en [47] se propone un método de entrenamiento en dos pasos, en el que en primer lugar se usa una red de entrenamiento facial para regularizar a nivel de característica las capas convolucionales de la red de reconocimiento emocional, forzando que la salida de ambas redes se aproxime, para posteriormente añadir una capa completamente conexa y entrenar la red completa usando las etiquetas emocionales.

Recientemente, el uso de *generative adversarial networks* (GANs) [48] también ha sido explorado en el estudio del reconocimiento de expresiones faciales invariante a pose o identidad, dada su gran utilidad en la síntesis de imágenes. En [49] se propone un entrenamiento multitarea basado en GANs para realizar una frontalización facial, preservando la información de identidad y de expresión facial. El generador de la red sintetiza imágenes frontalizadas a partir de las imágenes de entrada, a la vez que aprende a clasificar las expresiones faciales en estas nuevas representaciones, mientras que el discriminador por su parte trata de distinguir entre imágenes frontales reales y las sintetizadas por el generador. Con el objetivo de entrenar clasificadores multivista, en [50] se propone un modelo basado en GANs para generar imágenes con diferentes expresiones faciales en distintas poses. Finalmente, en [51] desarrollan un método invariante a identidad con dos ramas, en la que una de ellas genera imágenes del usuario de entrada variando su expresión facial, mientras que la otra se trata de un reconocedor de expresiones faciales, el cual mide la distancia entre las características de la imagen de entrada y las extraídas de las imágenes de sintetizadas. Como resultado, se obtiene un clasificador que selecciona la emoción con menor distancia a la imagen de entrada, reduciendo así la variación inter-usuario.

2.2. Estudio de sesgos e información sensible en los algoritmos de aprendizaje automático

Como ya se adelantó en la Sección 1.1, el estudio de la aparición, reproducción y compensación de sesgos en los algoritmos de aprendizaje automático e IA, así como su capacidad de extracción de información sensible (e.g. género, etnia, edad, etc.), ha ido despertando un gran interés a lo largo de los últimos años. De forma genérica, podemos definir el concepto de "sesgo" en el aprendizaje automático como la presencia sistemática de uno o más prejuicios debido a las asunciones que se han tomado en uno o más puntos del proceso de aprendizaje, y que puedan afectar a las decisiones de un clasificador resultante de dicho proceso. Así, el sesgo puede ser introducido en el conjunto de entrenamiento (e.g. en la distribución de los datos o en la función objetivo), en el preprocesado de los datos (e.g. en la selección de características), en la estrategia de aprendizaje (e.g en la función de pérdidas o en el modelo a optimizar), o incluso en la interpretación de los resultados (e.g. en el umbral de clasificación).

De todos los puntos en los que puede aparecer el sesgo, el conjunto de datos a usar durante el proceso de aprendizaje puede ser de los más críticos. Muchos trabajos han tratado con los problemas asociados al mismo, ya sea por la capacidad de generalización de los algoritmos entrenados [52], el desajuste entre los datos de entrenamiento y evaluación [53], o la correlación entre el sesgo y la potencial información sensible presente en los mismos. Este último punto es de especial interés ante la puesta en marcha de algoritmos automáticos en la sociedad, dadas las regulaciones internacionales[2] que ponen el foco en la privacidad de los usuarios y la lucha contra la discriminación algorítmica.

Dado que los datos usados en el diseño de estos sistemas son recolectados en nuestra sociedad, los modelos resultantes pueden reflejar sesgos actuales e históricos de esta [3], e incluso llegar a amplificarlos [54, 55]. Esta situación, así como el uso de la información sensible de los usuarios, es aún más difícil de combatir cuando trabajamos con fuentes de información no estructurada, como pueden ser las imágenes, el audio, o el texto. Aunque los sistemas no sean entrenados explícitamente para reconocer dicha información sensible en estos dominios, en muchas ocasiones los modelos basados en *deep learning* pueden explotar la correlación en los datos o capturar parte de esta información de forma residual, y por ende sus decisiones acabarán siendo influenciadas por esta.

Eliminar la información sensible por completo en la entrada de los sistemas es prácticamente inviable [56] en estas condiciones. A su vez, la recolección de grandes bases de datos que reflejen de forma precisa y balanceada la realidad que se pretende modelar puede ser extremadamente costoso. Siendo conscientes de estas limitaciones, gran parte del estudio en este ámbito se centra en el desarrollo de métodos que prevengan el acceso a información sensible y la discriminación algorítmica cuando trabajamos con conjuntos de datos de estas características.

[2] `https://gdpr.eu/`

2.2.1. Eliminación de sesgos e información sensible en los algoritmos de aprendizaje automático

A lo largo de esta sección se presentarán diversos métodos que han sido propuestos en los últimos años para preservar la privacidad de los usuarios y prevenir la aparición de sesgos. Antes de comenzar esta revisión, es conveniente aclarar la diferencia entre corregir sesgos y eliminar información sensible. La aparición de sesgos suele estar asociado con un problema en la representación de los distintos grupos del conjunto de datos (e.g. clases desbalanceadas) [52] o con un etiquetado de los mismos que incluya criterios subjetivos, lo que lleva a un rendimiento desigual entre grupos [57]. Por lo tanto, un algoritmo corrector de sesgos efectúa un aprendizaje que compense dicho sesgo y equilibre el rendimiento entre las distintas clases. Por su parte, la eliminación de información sensible suele trabajar en la línea de la generación de nuevas representaciones de los datos en las que no se pueda extraer un determinado tipo de información (e.g. género, etnia, identidad...), convirtiéndose en representaciones "ciegas" a dicho atributo. De esta forma, aunque en muchas ocasiones ambas ideas van asociadas, se podría corregir un sesgo sin eliminar la información sensible asociada al mismo, e incluso eliminar dicha información sin que ello conllevará una corrección de un posible sesgo asociado.

Con esto en mente, son varios los métodos que en los últimos años se han propuesto para llevar a cabo estas tareas, actuando en diversos puntos del proceso de aprendizaje. Por ejemplo, en lo que concierne a las reglas de decisión, varios trabajos han tratado de combatir la discriminación a partir de técnicas de *discrimination-aware data mining* [58, 59]. Otra aproximación es la propuesta en [60], donde se modifica la función de verosimilitud en modelos probabilísticos, introduciendo unas etiquetas objetivo latentes para asegurar que se cumpla un criterio de equidad prefijado.[3] Siguiendo con clasificadores probabilísticos, en [61] se propone un clasificador de Bayes modificado, en el que se ajusta la probabilidad condicionada de la salida respecto a un atributo protegido, con el objetivo de lograr un rendimiento similar entre grupos. Finalmente, en [62] se propone un framework en el que se permita a un administrador humano seleccionar sus preferencias de cara a cómo realizar el *trade-off* entre la utilidad y los criterios de equidad[3] en el diseño de un clasificador, ya que en muchas ocasiones su optimización simultánea no es posible. Una vez seleccionadas dichas preferencias, se calcula una ponderación para cada término, y se busca un modelo óptimo que maximice con dicha configuración todos los criterios, seleccionando un umbral de clasificación determinado para cada grupo demográfico del problema.

Frente a estos métodos, otros muchos ponen el foco en el dominio de entrada del algoritmo, llevando a cabo un procesado de los datos en la entrada que pueda prevenir la aparición de sesgos. En [56] se demuestra cómo la eliminación de indicadores explícitos de género en diferentes representaciones semánticas no basta para eliminar el sesgo de género de un clasificador de ocupación basado en NLP, funcionando otras palabras como "proxy" de dicho atributo sensible. Los autores de [63] proponen eliminar información sensible, a la vez que se logre cierta interpretabilidad y se mantenga información semántica, apren-

[3] Existen diversos criterios de equidad de grupo con los que evaluar los algoritmos de clasificación, como pueden ser *demographic parity* o *equality of opportunity*.

Figura 2.2: Aplicación del método de [63] en el dataset CelebA [66]. En este caso, la transformación de las imágenes fue guiada por los atributos de cada imagen definidos en CelebA, tratando de corregir un sesgo de género en la clasificación de atractivo (i.e. en la partición de entrenamiento, un 70 % de los sujetos atractivos son mujeres, frente a un 30 % de hombres).

diendo una transformación de los datos en el dominio de entrada (e.g. imágenes o tablas) a una representación en el mismo dominio que cumpla cierto criterio de equidad. Partiendo de la asunción de que el dominio de entrada puede descomponerse en una componente dependiente del atributo protegido y otra independiente, se entrena una transformación basada en *neural style transfer* y *reproducing kernel Hilbert spaces*, pudiendo observar un ejemplo visual en la Figura 2.2. Una aproximación similar es la propuesta en [64], en la que se busca generar un nuevo dataset similar a un dataset multimedia dado, pero más justo respecto a un atributo protegido en la toma de decisiones. Para ello, se hace uso de GANs auxiliares [65], a las que se les incluye un término de pérdidas que fuerce un criterio de equidad.

Un grueso importante del trabajo en este campo se centra en la estrategia de aprendizaje para generar clasificadores libres de sesgos e información sensible. Los autores de [67] proponen una adaptación de las *Domain Adversarial Neural Networks* [68], originalmente desarrolladas para realizar una adaptación de dominio en la que los dominios de origen y destino provienen de distribuciones similares, para forzar a las CNNs a tomar decisiones sin basarse en un concepto protegido o en información contextual (e.g. fondo de la escena). También basado en adaptaciones de dominio, en [69] se propone mejorar el reconocimiento facial reduciendo sesgos raciales mediante una adaptación de dominio no supervisada, desde un dominio de origen etiquetado (i.e. individuos caucásicos) a uno no etiquetado (i.e. individuos de otras etnias). Dicho método parte de un *clustering* que genera pseudo-etiquetas con las que realizar una pre-adaptación al dominio objetivo, mejorando la capacidad discriminativa de la red con una segunda adaptación basada en la información mutua. En [70] se propone un método para mitigar los sesgos en la clasificación de ocupación sin tener que hacer uso de atributos sensibles, sino reduciendo la correlación entre la predicción del clasificador y el nombre de cada individuo. Con el objeto de eliminar información sensible de las representaciones extraídas por una red neuronal pre-entrenada, en [71] se propone una extensión de *triplet-loss* [72], con la que se elimine de forma simultánea esta información sin perder rendimiento de la tarea principal.

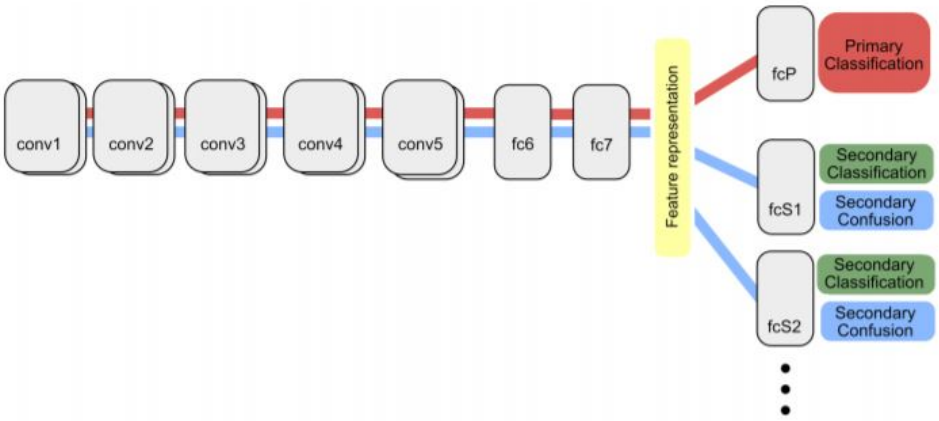


Figura 2.3: Arquitectura del método propuesto en [73], en la que una rama principal (fcP) entrena la tarea principal a partir de los *feature embeddings* extraídos, mientras que las ramas secundarias (fcSn) entrenan clasificadores de sesgos y métricas de confusión para eliminarlos de dichas representaciones.

Recientemente, se ha explorado el uso de estrategias de aprendizaje adversario [48] para prevenir el aprendizaje de sesgos y la extracción de información sensible. Por ejemplo, en [73] se desarrolla un método de aprendizaje y desaprendizaje conjunto (ver Figura 2.3), en el que simultáneamente una rama primaria aprende una representación útil para la tarea de clasificación principal, mientras que ramas secundarias de la red eliminan sesgos de dicha representación aplicando una función de confusión basada en la entropía entre la salida del mejor clasificador de dicho sesgo y una distribución uniforme. Los autores de [53] proponen una función de regularización en redes neuronales que minimice la información mutua entre los *feature embeddings* extraídos y el sesgo, usando una red adicional que predice la distribución del sesgo junto a técnicas de reversión del gradiente [68]. De forma similar, en [74] se busca generar representaciones que no puedan ser usadas de forma maliciosa por terceros, desarrollando funciones objetivo en esquemas de aprendizaje adversario que incluyan criterios de equidad de grupo para prevenir la extracción de información sensible.

Capítulo 3

Diseño del sistema

En este capítulo describiremos el sistema diseñado para aprender representaciones faciales emocionalmente ciegas, así como las herramientas seleccionadas para su desarrollo y posterior evaluación en el Capítulo 4. Más concretamente, en la Sección 3.1 se presenta la formulación del problema, para posteriormente introducir los dos métodos del Estado del Arte seleccionados para llevar a cabo dicha tarea, en las Secciones 3.1.1 y 3.1.2, y los detalles de implementación en la Sección 3.2. Finalmente, en la Sección 3.3 describiremos las distintas bases de datos utilizadas a lo largo de este trabajo.

3.1. Aprendizaje de representaciones faciales emocionalmente ciegas: Formulación del problema

Como ya se introdujo en la Sección 1.2, el objetivo principal de este trabajo es el aprendizaje de representaciones faciales emocionalmente ciegas, es decir, representaciones que sean útiles en alguna tarea relacionada con el análisis facial, pero que no puedan ser usadas en tareas de reconocimiento y/o análisis emocional.

Definiendo este problema formalmente, supongamos un sistema genérico orientado a una tarea k típica de análisis facial (e.g. reconocimiento de emociones, género o etnia) como el ilustrado en la Figura 3.1. Dicho sistema toma como entrada una imagen facial $\mathbf{I_x}$, de la que extrae una representación facial o *feature embedding* $\mathbf{x} \in \mathbb{R}^N$ mediante un modelo pre-entrenado de parámetros $\mathbf{w}$, al que nos referimos como extractor de características. Este modelo suele estar entrenado con grandes bases de datos en una tarea $k = 0$, próxima a la tarea objetivo, que le permita extraer información con alta capacidad discriminativa, siendo típico que el *embedding* $\mathbf{x}$ se obtenga a la salida de una de las últimas capas de una CNN. En nuestra arquitectura, partimos del reconocimiento facial como tarea $k = 0$.

Sobre estas representaciones se puede aplicar una transformación $\mathbf{f}_k(\mathbf{x})\,(k \geq 1)$ que realice una adaptación de dominio a una nueva tarea. En el sistema de la Figura 3.1, dichas transformaciones son implementadas mediante una capa densa con activación ReLU,

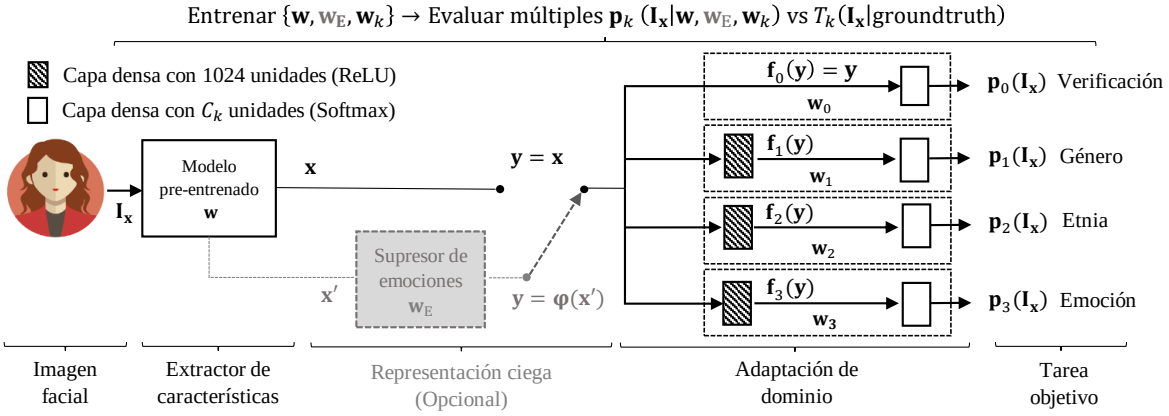


Figura 3.1: Esquema general del entorno experimental propuesto, incluyendo la adaptación de dominio desde un modelo pre-entrenado a varias tareas k, con o sin representaciones emocionalmente ciegas $\boldsymbol{\varphi}(\mathbf{x}')$. C_k es el número de clases de la tarea k, $\mathbf{f}_k$ es la transformación al dominio adaptado, y $\mathbf{p}_k$ es un vector con las probabilidades de pertenencia a cada clase de la tarea k.

añadiendo a continuación una segunda capa densa con activación *softmax* de C_k unidades, la cual realiza la predicción final. La salida en la rama de cada tarea es un vector $\mathbf{p}_k(\mathbf{I_x})$ de tamaño C_k con las probabilidades de que $\mathbf{I_x}$ pertenezca a cada una de las clases de la tarea k. A esta segunda parte del sistema, que incluye tanto la transformación $\mathbf{f}_k(\mathbf{x})$ como la capa *softmax*, nos referiremos como clasificador, el cual estará representado por sus parámetros $\mathbf{w}_k$.

El entrenamiento de cada tarea consiste, por lo tanto, en encontrar los parámetros $\mathbf{w}_k$, y opcionalmente $\mathbf{w}$ si no se deja fijo el extractor de características, que minimicen el error entre la salida $\mathbf{p}_k(\mathbf{I_x})$ y una función objetivo $T_k(\mathbf{I_x})$. Partiendo del conocimiento del *groundtruth*, la función objetivo T_k indica la salida esperada para cada entrada $\mathbf{I_x}$ (e.g. en la tarea de reconocimiento de género $k = 1$, $T_1 = 0$ para la clase hombre, $T_1 = 1$ para la clase mujer). La estrategia de aprendizaje más común consiste en minimizar una función de pérdidas $\mathcal{L}$, siendo común el uso de la función *cross-entropy*, sobre un conjunto de muestras $\mathcal{P}$ para las que tenemos un *groundtruth*:

$$\min_{\mathbf{w},\mathbf{w}_k} \sum_{\mathbf{I_x}\in\mathcal{P}} \mathcal{L}[\,\mathbf{p}_k(\mathbf{I_x}|\mathbf{w},\mathbf{w}_k)\,,\,T_k(\mathbf{I_x})\,] \tag{3.1}$$

Con este esquema en mente, recordemos que el objetivo de este trabajo es la generación de representaciones faciales emocionalmente ciegas, es decir, buscamos generar una representación con la que se pueda entrenar una tarea $k \neq 3$ distinta al reconocimiento de emociones. Para poder alcanzar dicho objetivo, al esquema descrito previamente se le añade un módulo supresor de información emocional (ver Figura 3.1) de parámetros $\mathbf{w}_\mathrm{E}$ que, partiendo de la información $\mathbf{x}'$ que recibe del extractor de características (que podría ser el mismo *embedding* $\mathbf{x}$ en función de la implementación), genera una representación $\boldsymbol{\varphi}(\mathbf{x}') \in \mathbb{R}^M$ carente de información emocional. Teniendo esto en cuenta, la estrategia de aprendizaje presentada en la Ecuación 3.1 se debe modificar añadiendo un nuevo término al objetivo de optimización:

$$\min_{\mathbf{w},\mathbf{w}_k,\mathbf{w}_\mathrm{E}} \sum_{\mathbf{I_x}\in\mathcal{S}} \{\mathcal{L}[\,\mathbf{p}_k(\,\mathbf{I_x}|\mathbf{w},\mathbf{w}_\mathrm{E},\mathbf{w}_k)\,,\,T_k(\,\mathbf{I_x})\,] + \lambda\,\mathcal{L}_r[\,\mathbf{p}_3(\,\mathbf{I_x}|\mathbf{w},\mathbf{w}_\mathrm{E},\mathbf{w}_k)\,,\,T_3(\,\mathbf{I_x})]\} \tag{3.2}$$

En esta nueva estrategia, $\mathcal{L}_r$ representa una función de pérdidas orientada a reducir el rendimiento en la tarea de reconocimiento de emociones $k = 3$, mientras que la función $\mathcal{L}$ trata de maximizar el rendimiento en una tarea $k \neq 3$, y λ es un hiper-parámetro que balancea ambos términos. Cabe destacar que, en este caso, el conjunto de datos usados en el entrenamiento $\mathcal{S}$ puede ser diferente al $\mathcal{P}$ usado previamente, de forma que se incluyan muestras con información emocional que ayuden a eliminar esta información en el entrenamiento.

Con el objetivo de desarrollar distintas implementaciones del supresor de emociones, en este trabajo proponemos el uso de dos métodos del Estado del Arte en corrección de sesgos y eliminación de información sensible. Estos métodos, conocidos como Sensitive-Nets (SN) [71] y Learning Not To Learn (LNTL) [53], ya se introdujeron brevemente en la Sección 2.2.1, pudiendo encontrar en las siguientes secciones una descripción más detallada de los mismos.

3.1.1. SensitiveNets

SensitiveNets [71] es un método propuesto recientemente para eliminar información de género y etnia de los *embeddings* extraídos por un modelo pre-entrenado, sin provocar una caída significativa en el rendimiento de estos en el reconocimiento facial. Para aplicar este método al sistema presentado en la Figura 3.1, dejamos fijos los parámetros $\mathbf{w}$ del extractor de características, y optimizamos la siguiente versión de la Ecuación 3.2 a partir de las representaciones extraídas por este:

$$\min_{\mathbf{w}_\mathrm{E}} \sum_{\mathrm{triplet}\in\mathcal{S}_p} \{\mathcal{L}_t[\boldsymbol{\varphi}_\mathrm{SN}(\mathbf{x}_\mathrm{triplet}|\mathbf{w},\mathbf{w}_\mathrm{E})] + \Delta^\mathrm{A} + \Delta^\mathrm{P} + \Delta^\mathrm{N}\} \tag{3.3}$$

$$\mathcal{L}_t = \|\boldsymbol{\varphi}_\mathrm{SN}(\mathbf{x_A}) - \boldsymbol{\varphi}_\mathrm{SN}(\mathbf{x_P})\|^2 - \|\boldsymbol{\varphi}_\mathrm{SN}(\mathbf{x_A}) - \boldsymbol{\varphi}_\mathrm{SN}(\mathbf{x_N})\|^2 + \alpha \tag{3.4}$$

En la Ecuación 3.3, los triplet $= \{\mathbf{I_A}, \mathbf{I_P}, \mathbf{I_N}\}$ están formados por dos imágenes de un mismo usuario (i.e. $\mathbf{I_A}$ e $\mathbf{I_P}$), y una de un usuario diferente (i.e. $\mathbf{I_N}$), $\mathcal{L}_t$ es la función de pérdidas *triplet-loss* [72] con umbral α, y los términos Δ son regularizadores que miden la información emocional en las representaciones del supresor emocional $\mathbf{w}_\mathrm{E}$:

$$\Delta = \log\{\,1 + |\,0{,}9 - P_3(\mathrm{Neutral}\,|\boldsymbol{\varphi}_\mathrm{SN}(\mathbf{x}), \mathbf{w}_3)\,|\} \tag{3.5}$$

La probabilidad P_3 de observar una expresión neutra en los *embeddings* tras pasar por el supresor de emociones es obtenida usando un clasificador de emociones $\mathbf{w}_3$ entrenado

con las representaciones $\boldsymbol{\varphi}_{\mathrm{SN}}(\mathbf{x})$, utilizando un conjunto $\mathcal{S}_E$ que puede ser distinto al usado en la Ecuación 3.3. Cuando esta probabilidad toma valores altos, los términos regularizadores tienden a cero. Por lo tanto, al minimizar la Ecuación 3.3 durante el entrenamiento, estaremos forzando al supresor de emociones a que las representaciones extraídas sean en general expresiones neutras, eliminando las características necesarias para distinguir entre emociones.

El aprendizaje de SN se aplica siguiendo un esquema adversario [48], en primer lugar minimizando la información emocional en las representaciones $\boldsymbol{\varphi}_{\mathrm{SN}}(\mathbf{x})$ mientras que se mantiene el rendimiento en reconocimiento facial gracias a la función *triplet-loss*, para posteriormente entrenar un nuevo clasificador emocional con estas, de forma que se adapte a las nuevas características y extraiga toda la información emocional posible. Iterando en este proceso, se acaba obteniendo una representación facial que mantiene su utilidad, pero carece de información emocional.

3.1.2. Learning Not To Learn

Learning Not To Learn [53] es un método de entrenamiento de DNNs, en el que se propone un término de regularización para prevenir que la red aprenda un sesgo conocido en los datos que sea irrelevante o indeseado en la tarea principal. Aunque no se trate de un sesgo, en este trabajo proponemos usar este método para eliminar la información emocional del sistema de la Figura 3.1.

Este algoritmo usa una función de pérdidas que incluye un término de regularización, basado en la información mutua entre el sesgo que se desea eliminar, y los *embeddings* $\boldsymbol{\varphi}_{\mathrm{LNTL}}(\mathbf{x}')$. Dichos *embeddings* son utilizados posteriormente en la tarea principal $k \neq 3$, cuyo rendimiento es maximizado mediante una función *cross-entropy*. Juntando todos los elementos, el objetivo de optimización de este método es el siguiente:

$$\min_{\mathbf{w}_{\mathrm{E}},\mathbf{w}_k} \sum_{\mathbf{I}_\mathbf{x}\in\mathcal{S}} \{\mathcal{L}_c[\,\mathbf{p}_k(\mathbf{I}_\mathbf{x}|\mathbf{w},\mathbf{w}_{\mathrm{E}},\mathbf{w}_k)\,,\,T_k(\mathbf{I}_\mathbf{x})\,] + \lambda\,\mathcal{I}[\,T_3(\mathbf{I}_\mathbf{x})\,;\,\boldsymbol{\varphi}_{\mathrm{LNTL}}(\mathbf{x}')\,]\,\} \tag{3.6}$$

$$\mathcal{I}[\,T_3(\mathbf{I}_\mathbf{x})\,;\,\boldsymbol{\varphi}_{\mathrm{LNTL}}(\mathbf{x}')\,] = H[\,T_3(\mathbf{I}_\mathbf{x})] - H[\,T_3(\mathbf{I}_\mathbf{x})|\,\boldsymbol{\varphi}_{\mathrm{LNTL}}(\mathbf{x}')] \tag{3.7}$$

En la Ecuación 3.6, $\mathcal{I}[\,;\,]$ representa la información mutua, y λ es un hiperparámetro que balancea ambos términos. Si prestamos atención a la expresión de la información mutua (ver Ecuación 3.7), podemos observar que la entropía marginal es una constante independiente de los parámetros de entrenamiento, y por lo tanto sólo es necesario optimizar la entropía condicionada negativa $-H[\,T_3(\mathbf{I}_\mathbf{x})|\,\boldsymbol{\varphi}_{\mathrm{LNTL}}(\mathbf{x}')]$. Dado que para calcular esta entropía es necesario conocer la distribución a posteriori del sesgo dados los *embeddings*, en la práctica se usa una red $\mathbf{w}_3$ entrenada a partir de estas representaciones en la tarea de reconocimiento de emociones $k = 3$ para parametrizar dicha distribución,

asumiendo que al ser entrenada con el *groundtruth* $T_3(\mathbf{I_x})$ su salida la aproximará. Por lo tanto, el objetivo de optimización final será:

$$\min_{\mathbf{w}_\mathrm{E},\mathbf{w}_k} \max_{\mathbf{w}_3} \sum_{\mathbf{I_x}\in\mathcal{S}} \{\mathcal{L}_c[\,\mathbf{p}_k(\mathbf{I_x}|\mathbf{w},\mathbf{w}_\mathrm{E},\mathbf{w}_k)\,,\,T_k(\mathbf{I_x})\,] - \lambda\, H[\,\mathbf{p}_3(\mathbf{I_x}|\mathbf{w},\mathbf{w}_\mathrm{E},\mathbf{w}_3)|\,\boldsymbol{\varphi}_\mathrm{LNTL}(\mathbf{x}')] - \mathcal{L}_c[\,\mathbf{p}_3(\mathbf{I_x}|\mathbf{w},\mathbf{w}_\mathrm{E},\mathbf{w}_3)\,,\,T_3(\mathbf{I_x})\,] \tag{3.8}$$

En la Ecuación 3.8 se puede apreciar la estrategia adversaria [48] que sigue este método. El clasificador $\mathbf{w}_3$ se entrena para discriminar correctamente entre emociones faciales. Por su parte, el supresor de emociones $\mathbf{w}_\mathrm{E}$ busca extraer características que dificulten el reconocimiento de emociones, mediante la minimización de la entropía condicionada negativa, pero que sean útiles en la tarea k que entrena el clasificador primario. En la práctica, además de la estrategia adversaria se utiliza la técnica de inversión de gradiente [68]

3.2. Detalles de implementación

El punto de partida fundamental en nuestra implementación es definir concretamente la tarea de reconocimiento de emociones a partir de expresiones faciales que consideraremos en este trabajo. Como vimos en la Sección 2.1, las emociones suelen asociarse a expresiones faciales prototípicas, que podemos describir mediante una combinación de movimientos faciales o *action units*. Esto nos lleva a plantearnos que, si queremos evitar que se pueda explotar información emocional en los *embeddings* diseñados, se podría suprimir la información referente a dichas AUs de estos, eliminando por lo tanto la capacidad de reconocer expresiones faciales. Sin embargo, esta aproximación plantea un problema complejo, puesto que el reconocimiento de AUs es una tarea inherentemente multi-etiqueta, lo cual complica significativamente la eliminación de esta información con los métodos propuestos.

Para evitar esta complicación, trabajaremos directamente con expresiones faciales asociadas a las emociones básicas, a saber: Neutra, Felicidad, Tristeza, Enfado, Sorpresa y Disgusto. Por lo tanto, estaremos trabajando con un problema multi-clase de 6 posibles clases. La expresión facial asociada al Miedo se ha dejado fuera del desarrollo de este trabajo, dado que en los primeros experimentos que llevamos a cabo las muestras correspondientes a esta emoción eran escasas o ruidosas, por lo que se decidió prescindir de esta en adelante.

Volviendo de nuevo al esquema general presentado en la Figura 3.1, en este trabajo hemos usado el modelo pre-entrenado ResNet-50 [75] como extractor de características $\mathbf{w}$ con el que generar las representaciones faciales. ResNet-50 es una CNN muy popular en los campos del reconocimiento facial y del análisis de imagen, compuesta por 50 capas que incluyen conexiones "residuales". Estas conexiones permiten a la red combinar la salida de una capa con la entrada de capas previas, implementando un *identity mapping* que ayude en la propagación de los gradientes en el entrenamiento y mejore el rendimiento de la red

según aumenta su profundidad (i.e. se soluciona el problema conocido como *vanishing gradients* [75]), a la vez que reduce significativamente el número de parámetros). Más concretamente, hemos usado la versión de ResNet-50 tuneada con VGG2 [76] para la tarea de reconocimiento facial, siendo una red más adecuada al problema en cuestión. La última capa convolucional de esta red será la salida de nuestro extractor de características, siendo dicha salida *face embeddings* de 2048 características.

Dado que ResNet-50 acepta como entrada imágenes faciales de 224 x 224 píxeles, realizamos un preprocesado sobre todas las imágenes usadas en este trabajo, consistente en una detección facial seguida de un escalado. Con este fin, usamos el detector facial MTCNN [44] para extraer *bounding boxes* que delimiten la cara y con las que recortar las imágenes, para posteriormente reescalar las imágenes a 224 x 224 píxeles.

Respecto al supresor de emociones $\mathbf{w}_{\mathrm{E}}$, en el caso de SensitiveNets [71] se ha implementado usando una capa densa de 1024 unidades y activación lineal, congelando completamente el extractor. Por su parte, al aplicar Learning Not To Learn [53] decidimos no añadir nuevas capas al extractor, sino que usamos este método directamente sobre ResNet, dejando las última capas sin congelar durante el entrenamiento. Por lo tanto, los embeddings $\boldsymbol{\varphi}_{\mathrm{SN}}(\mathbf{x})$ tendrán una dimensionalidad de 1024, frente a las 2048 características de $\boldsymbol{\varphi}_{\mathrm{LNTL}}(\mathbf{x}')$.

Finalmente, en lo concerniente a los clasificadores de las distintas tareas, salvo en el caso del reconocimiento facial todos ellos se han diseñado como una capa densa de 1024 unidades y activación ReLU, seguida de de una capa con activación *softmax* y C_k unidades. Dada la naturaleza del del reconocimiento facial, y siendo esta la tarea original del extractor de características, se usará directamente la distancia euclídea entre los *embeddings* extraídos para realizar las comparaciones genuinas e impostoras (i.e.verificación facial).

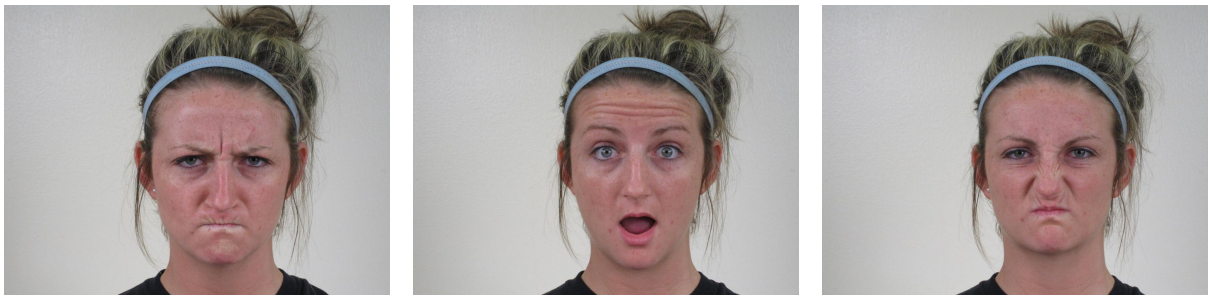
Todo el trabajo se ha desarrollado en el lenguaje de programación Python, usando la distribución de Anaconda. Más concretamente, nuestro sistema se ha desarrollado usando librerías comunes como Numpy, Pandas o Scikit-Learn, y Keras/TensorFlow como framework de *deep learning*, salvo en la implementación de LNTL que se realizó en PyTorch. La implementación de SN se ha realizado partiendo del código facilitado directamente por los autores, mientras que LNTL parte del código facilitado en su repositorio en GitHub[1] y la versión para PyTorch de ResNet-50[2] tuneada con VGG2.

3.3. Bases de datos

A lo largo de este trabajo hemos hecho uso de diversas bases de datos en nuestros experimentos, ya sea para entrenar los algoritmos como para su posterior evaluación. Estos datasets son los siguientes:

[1] `https://github.com/feidfoe/learning-not-to-learn`
[2] `https://github.com/ox-vgg/vgg_face2`

(a) Neutra (b) Felicidad (c) Tristeza



(d) Enfado (e) Sorpresa (f) Disgusto

Figura 3.2: Ejemplos visuales de las expresiones faciales asociadas a las emociones usadas en este trabajo, extraídas de la base de datos CFEE [25].

- CelebA [66]: El dataset CelebA[3] contiene 202,599 imágenes faciales de celebridades, contando con un total de 10,177 identidades distintas. Cada imagen tiene asociadas como anotaciones 5 puntos de referencia faciales, así como 40 atributos binarios entre los que se incluyen etiquetas de apariencia, género, edad, atractivo o estado emocional. Un ejemplo visual con varias imágenes de este dataset se presentó en la Figura 2.2.

- DiveFace [71]: La base de datos DiveFace[4] contiene 24,000 identidades distintas, distribuidas de forma equitativa en 6 grupos demográficos en relación a etnia y género. Con más de 120,000 imágenes, cada individuo se asigna a una clase en función de su género (i.e. hombre o mujer) y sus rasgos étnicos, clasificados en 3 grupos étnicos distintos, a saber: Asia Oriental, Caucásico, y África subsahariana e India.

- CFEE [25]: El dataset Compound Facial Expressions of Emotion[5] incluye imágenes faciales de 230 individuos, tomadas en condiciones controladas de laboratorio. Cada individuo presenta una imagen por cada una de las 22 categorías del dataset, a saber: expresión neutral, 6 emociones básicas y 15 emociones compuestas. Todas las imágenes representan de forma inequívoca una expresión facial reconocible. Además, durante la realización de este trabajo etiquetamos manualmente la base de datos con anotaciones de género. Un ejemplo visual de estas imágenes se presenta en la Figura 3.2.

[3] `http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`
[4] `https://github.com/BiDAlab/DiveFace`
[5] `http://cbcsl.ece.ohio-state.edu/dbform_compound.html`

- LFW [77]: Labeled Faces in the Wild[6] es un conocido dataset y *benchmark* público para el estudio y evaluación de algoritmos de reconocimiento facial. Contiene más de 13,000 imágenes, y 5,749 identidades, recolectadas de la web. En este trabajo hemos hecho uso de las imágenes de test alineadas [78], y su protocolo de evaluación asociado en la vista 1.

[6] `http://vis-www.cs.umass.edu/lfw/`

Capítulo 4

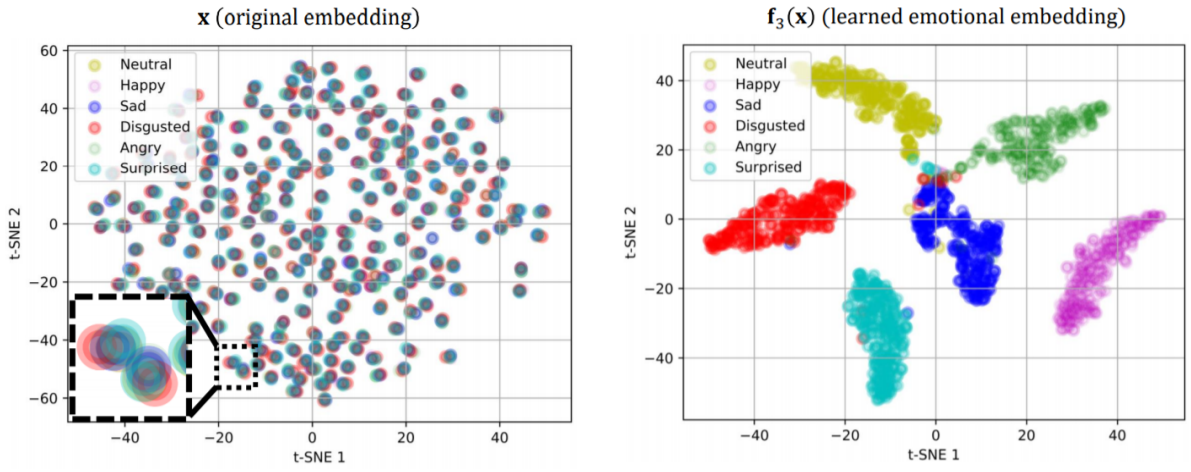
Experimentos y Resultados

En este capítulo describiremos los experimentos realizados en este trabajo, a la vez que presentaremos y analizaremos los distintos resultados obtenidos. Nuestros experimentos se pueden dividir en 3 bloques distintos y consecuentes, abordando en distintas secciones cada uno de ellos. En la Sección 4.1 estudiaremos la información emocional presente en las representaciones faciales extraídas por ResNet-50 [75], la CNN de reconocimiento facial que escogimos para la realización de este trabajo. A partir de de los resultados de dicho estudio, en la Sección 4.2 nos centramos en eliminar dicha información aplicando los métodos presentados en las Secciones 3.1.1 y 3.1.2. Finalmente, en la Sección 4.3 se plantea un ejemplo de aplicación de las nuevas representaciones faciales en la corrección de sesgos.

4.1. Estudio de la información emocional codificada en representaciones faciales

Con el objetivo de eliminar la información referente a las expresiones faciales de las representaciones extraídas por las redes neuronales profundas, el primer paso lógico es estudiar cómo está codificada esta información, y hasta qué punto puede ser explotada. Recordemos que, en el marco de este trabajo, entendemos que se puede "explotar" esta información si a partir de las representaciones faciales se puede entrenar un clasificador de utilidad, entendiendo dicha utilidad como un rendimiento significativamente superior al clasificador aleatorio (i.e. para un problema de clasificación en 6 clases, una precisión del 16,67 %). Para ello, en esta sección haremos uso de la base de datos CFEE [25], previamente introducida en la Sección 3.3, para explorar la presencia de esta información en los *embeddings* $\mathbf{x}$ generados por ResNet-50 [75].

En la Figura 4.1(a) se muestran las proyecciones bidimensionales de los *embeddings* extraídos de CFEE, usando el algoritmo de reducción de dimensionalidad t-SNE y etiquetando los datos posteriormente para facilitar su interpretación. Esta representación nos permite visualizar datos de muy alta dimensionalidad en un espacio de 2 o 3 dimensiones,



(a) (b)

Figura 4.1: Proyecciones t-SNE de las representaciones faciales originales generadas por ResNet-50 [75] (a), y las adaptadas al dominio de reconocimiento de emociones (b), extraídas de las imágenes de CFEE [25].

de forma que objetos similares se modelan como puntos cercanos en el nuevo espacio con alta probabilidad. Como puede observarse, las representaciones faciales ignoran la información emocional, formando pequeñas agrupaciones en las que se solapan muestras de cada una de las expresiones. Si hacemos memoria, el modelo de ResNet-50 que estamos utilizando ha sido entrenado para llevar a cabo tareas de reconocimiento facial. Esto nos lleva a que en los *embeddings* extraídos la información predominante sea discriminativa respecto a la identidad, por lo que al proyectarlos usando t-SNE se forman pequeñas agrupaciones que pueden interpretarse como los distintos usuarios de CFEE. En este sentido, la información referente a la expresión facial puede entenderse como una variación espuria de poca utilidad para identificar a los usuarios (i.e. la identidad de un usuario es independiente de la expresión facial que esté demostrando), por lo que no debería estar incluida en una representación entrenada para el reconocimiento facial.

Sin embargo, que no sea la información predominante en las representaciones faciales no significa que no exista información latente sobre la expresión facial. Para poder revelar dicha información, a partir de los *embeddings* $\mathbf{x}$ entrenamos un clasificador de emociones $\mathbf{w}_3$. Este clasificador realiza una adaptación de dominio $\mathbf{f}_3(\mathbf{x})$ a un espacio óptimo para discriminar entre expresiones faciales. En la Figura 4.1(b) se muestran las proyecciones t-SNE de los vectores $\mathbf{f}_3(\mathbf{x})$ extraídos de CFEE. Tras el entrenamiento, las muestras han pasado a agruparse en base a la expresión facial representada, lo que indica la presencia en el *embedding* original de información suficiente para discriminar en base a la expresión facial, con la cual se puede obtener más de un 80 % de rendimiento en dicha tarea.

Ahora que hemos visto que la información emocional está presente en los *embeddings* $\mathbf{x}$, nos preguntamos en qué medida se encuentra codificada en estos. Con este fin, evaluamos el rendimiento del clasificador de emociones variando el número de características disponibles del *embedding* $\mathbf{x}$. En cada iteración, seleccionamos aleatoriamente un porcentaje determinado de características con las que entrenar $\mathbf{w}_3$, dejando en todo momento los

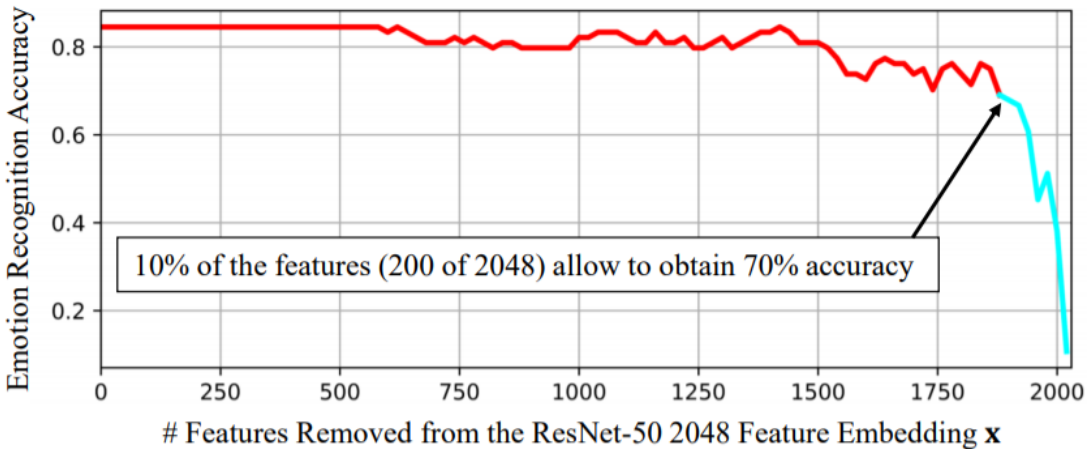


Figura 4.2: Rendimiento del clasificador de emociones $\mathbf{w}_3$ en función de las características de $\mathbf{x}$ disponibles durante su entrenamiento.

pesos del modelo ResNet. En la Figura 4.2 se presenta el rendimiento del clasificador en función el número de características disponibles durante el entrenamiento. Como puede observarse, el modelo es capaz de mantener un rendimiento superior al 80 % en precisión hasta el momento en el que se eliminan 1500 características (i.e. en torno al 70 %) de los *embeddings*, teniendo que eliminar hasta un 90 % de estas para lograr que decaiga hasta el 70 %. Esto demuestra que la información correspondiente a las expresiones faciales está latente en casi todas las características extraídas, en vez de tratarse de una información residual que pueda eliminarse mediante la supresión de un determinado número de estas.

4.2. Aprendizaje de representaciones faciales emocionalmente ciegas

Ahora que hemos comprobado que la información referente a las expresiones faciales está latente en los *embeddings* de ResNet [75], en esta sección buscamos eliminar dicha información, sin que se produzca una caída de rendimiento significativa en otras tareas de análisis facial (ver Figura 3.1). Con este fin, haremos uso de los métodos presentados en las Secciones 3.1.1 y 3.1.2, así como de las bases de datos introducidas en la Sección 3.3, para aprender las nuevas representaciones faciales emocionalmente ciegas.

En primer lugar, llevamos a cabo el entrenamiento de las representaciones $\boldsymbol{\varphi}_{\mathrm{SN}}(\mathbf{x})$ aplicando SensitiveNets [71] sobre las representaciones originales de ResNet (consultar Sección 3.2 para más detalles). Al aplicar este método mantenemos el rendimiento en una tarea principal con el dataset $\mathcal{S}_P$, mientras que se elimina la información emocional entrenando iterativamente un clasificador emocional $\mathbf{w}_3$ con un conjunto $\mathcal{S}_E$. En este caso hacemos uso de DiveFace [71] como conjunto $\mathcal{S}_P$, con el objetivo de mantener el rendimiento en verificación facial intacto tras eliminar la información emocional. Junto a este dataset, usamos la base de datos CFEE [25] como conjunto $\mathcal{S}_E$ con el que suprimir esta información.

Tarea	$\mathbf{x}$	$\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$	**Dif. SN**	$\boldsymbol{\varphi}_{\text{LNTL}}(\mathbf{x}')$	**Dif. LNTL**
ID	96,8 %	96,3 %	↓ 1 %	59,4 %	↓ 75,0 %
Género	99,2 %	98,9 %	↓ 1 %	72,7 %	↓ 53,9 %
Etnia	98,8 %	98,6 %	↓ 1 %	67,4 %	↓ 47,9 %
Emoción (NN)	88,1 %	59,6 %	↓ 40 %	41,6 %	↓ 65,0 %
Emoción (SVM)	88,1 %	16,7 %	↓ 100 %	25,0 %	↓ 88,2 %
Emoción (RF)	77,4 %	58,3 %	↓ 31 %	44,7 %	↓ 53,8 %

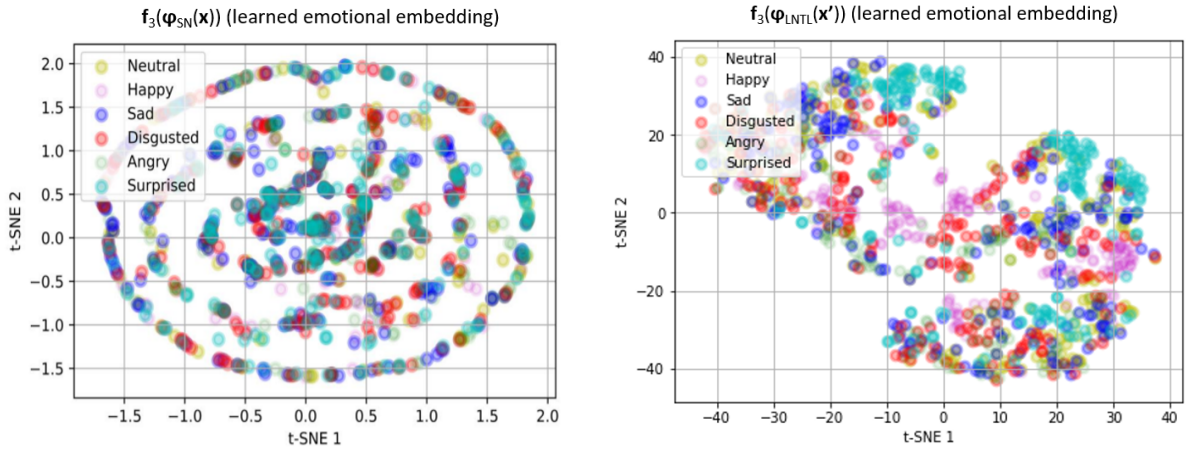
Tabla 4.1: Rendimiento de diferentes clasificadores entrenados tanto con las representaciones originales como con las emocionalmente ciegas.

Por otra parte, generamos una segunda versión de los *embeddings* emocionalmente ciegos haciendo uso de Learning Not To Learn [53]. A diferencia de SN, este método requiere que los conjuntos $\mathcal{S}_P$ y $\mathcal{S}_E$ sean el mismo durante el entrenamiento, por lo que debemos seleccionar una base de datos que incluya tanto anotaciones emocionales como de otra tara de análisis facial diferente. Para poder disponer de toda esta información, seleccionamos de nuevo la base de datos CFEE, y llevamos a cabo un etiquetado manual de la misma en cuanto a género. Por lo tanto, entrenamos las representaciones $\boldsymbol{\varphi}_{\text{LNTL}}(\mathbf{x}')$ en la tarea principal de reconocimiento de género ($k = 1$).

Una vez entrenamos los modelos con los que extraer los *embeddings* $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ y $\boldsymbol{\varphi}_{\text{LNTL}}(\mathbf{x}')$, realizamos una evaluación de los mismos con la que comprobar si se cumplen nuestros dos objetivos, a saber: (1) mantener la utilidad de los *embeddings* en diversas tareas de análisis facial, y (2) eliminar la información emocional. La Tabla 4.1 recoge los resultados de la evaluación de ambos objetivos en términos de precisión. Durante esta evaluación comprobamos la utilidad de las representaciones faciales en las distintas tareas presentadas en este trabajo, haciendo uso del dataset LFW [77] para comprobar el rendimiento en verificación facial, y de DiveFace para el reconocimiento de género y etnia. En el caso del reconocimiento de emociones, usamos CFEE para entrenar 3 algoritmos de clasificación distintos con NN, SVM y *Random Forest* (RF). En todos los casos, la caída de rendimiento se calcula siguiendo la Ecuación 4.1, donde Acc_k simboliza la precición en la tarea k, y Random_k es la precisión de un clasificador aleatorio en dicha tarea (i.e. 50 % en verificación facial y reconocimiento de género, 33,3 % en etnia, y 16,7 % en emociones).

$$\text{Dif}_k = (\text{Acc}_k(\mathbf{x}) - \text{Acc}_k(\boldsymbol{\varphi}))/(\text{Acc}_k(\mathbf{x}) - \text{Random}_k) * 100 \tag{4.1}$$

Si atendemos a nuestro primer objetivo, se puede observar una caída muy ligera de rendimiento en las tareas principales cuando usamos las representaciones $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$, lo que demuestra la eficacia de este método en preservar la información original. Por su parte, al usar $\boldsymbol{\varphi}_{\text{LNTL}}(\mathbf{x}')$, entrenados en la tarea principal de reconocimiento de género, la caída de rendimiento es considerablemente mayor en las 3 tareas respecto a las representaciones originales $\mathbf{x}$. La caída en verificación facial puede deberse a que LNTL es un método orientado a problemas con un bajo número de clases, mientras que el reconocimiento facial requiere un diseño orientado a un gran número de clases (i.e. una clase por identi-



(a) (b)

Figura 4.3: Proyecciones t-SNE de las representaciones faciales generadas con Sensitive-Nets [71] (a) y Learning Not To Learn [53] (b), extraídas de las imágenes de CFEE [25].

dad). Respecto a la caída en etnia y género, es posible que se deba a la diferencia entre CFEE y DiveFace, al estar compuesta la primera por imágenes tomadas en un entorno extremadamente controlado y con baja representatividad étnica, o a la relación entre la información eliminada sobre las expresiones faciales, y la propia de género y etnia

Centrándonos ahora en nuestro segundo objetivo, la reducción de rendimiento en el reconocimiento de emociones, se puede observar que ambos métodos logran una caída de rendimiento importante respecto a los *embeddings* $\mathbf{x}$ en los 3 algoritmos de clasificación. En ambos casos, la mayor caída de rendimiento se produce en los clasificadores entrenados con SVM, llegando al rendimiento de un clasificador aleatorio al usar $\boldsymbol{\varphi}_{\mathrm{SN}}(\mathbf{x})$ (i.e. 16,67 %), mientras que los clasificadores entrenados con RF mantienen un mayor rendimiento. Como vimos en la Sección 4.1, la información emocional está fuertemente embebida en las representaciones $\mathbf{x}$, y por lo tanto es difícil eliminarla por completo sin que se produzca una caída importante de rendimiento en otras tareas.

En la Figura 4.3 se pueden observar las proyecciones t-SNE similares a las presentadas en la Figura 4.1(b), pero calculadas a partir de las representaciones emocionalmente ciegas $\mathbf{f}_3(\boldsymbol{\varphi}_{\mathrm{SN}}(\mathbf{x}))$ $\mathbf{f}_3(\boldsymbol{\varphi}_{\mathrm{LNTL}}(\mathbf{x}'))$. A diferencia del caso con las representaciones originales, tras la adaptación de dominio el clasificador $\mathbf{w}_3$ sigue sin poder discriminar correctamente entre las distintas expresiones presentes en los datos. De hecho, en las proyecciones de LNTL se pueden apreciar dos grupos casi separados, que no se corresponden con grupos emocionales. En la Figura 4.4 mostramos esta misma proyección, pero esta vez etiquetando posteriormente los datos con las anotaciones de género, comprobando que las agrupaciones se han realizado de acuerdo a este. A pesar de entrenar dicha proyección en la tarea de reconocimiento de emociones, la poca información restante en los *embeddings* $\boldsymbol{\varphi}_{\mathrm{LNTL}}(\mathbf{x}')$ lleva a que la información de género siga siendo predominante en estos.

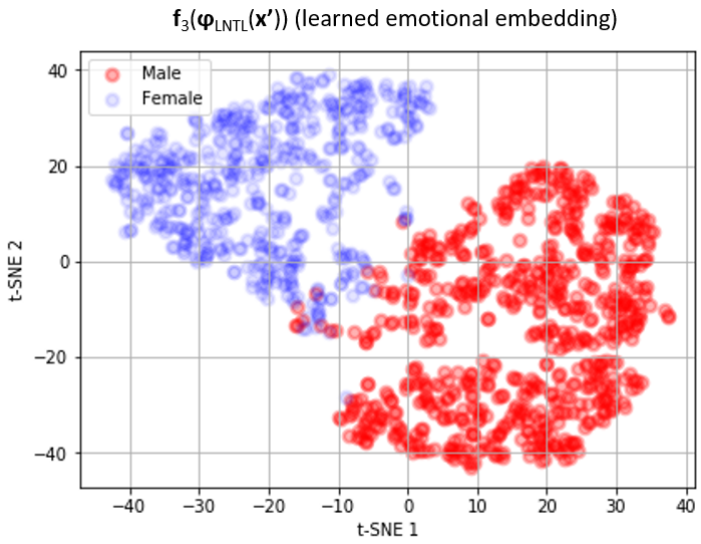


Figura 4.4: Proyecciones t-SNE de las representaciones faciales generadas con Learning Not To Learn [53], extraídas de las imágenes de CFEE [25] y etiquetadas por género.

4.3. Corrección de sesgos mediante el uso de representaciones faciales emocionalmente ciegas

En la Sección 4.1 demostramos que la información emocional está presente en las representaciones faciales extraídas por las redes de reconocimiento facial, para posteriormente comprobar en la Sección 4.2 que dicha información puede eliminarse sin que esto supongo una caída de rendimiento significativa en otras tareas de análisis facial. Ahora que hemos obtenido representaciones faciales en las que la información correspondiente a las emociones se ha suprimido, en esta sección desarrollaremos un experimento con el que ilustrar la utilidad de estas en la corrección de sesgos. Más concretamente, nos hemos inspirado en el experimento de [63] sobre sesgos de género para estudiar cómo las nuevas representaciones emocionalmente ciegas nos pueden ayudar a alcanzar un criterio de equidad específico.

En este experimento definimos una nueva tarea de análisis facial $k = 4$, la clasificación del atractivo a partir de una imagen de entrada $\mathbf{I_x}$, donde el objetivo será predecir si una persona es atractiva o no (i.e. $T_4(\mathbf{I_x}) = 1$ para la clase atractiva, $T_4(\mathbf{I_x}) = 0$ en caso contrario). La finalidad del experimento será estudiar si, mediante el uso de representaciones emocionalmente ciegas, se puede alcanzar el criterio de equidad *equality of opportunity* en presencia de un sesgo relacionado con un atributo protegido s. Este criterio es de gran utilidad en tareas de clasificación en las que la clase positiva $T_k = 1$ está asociada a un resultado ventajoso (e.g. obtención de crédito, contratación, etc.), requiriendo que la salida del clasificador $\mathbf{w}_k$ sea independiente del atributo s, condicionada a la clase positiva. La Ecuación 4.2 ilustra dicha condición aplicada a nuestro experimento, que en nuestro caso se traduce en obtener la misma tasa de verdaderos positivos (TPR) para los distintos valores de s por tratarse de un problema de clasificación binario. Nótese que

se han omitido las dependencias de $\mathbf{p}_4(\mathbf{I_x})$ a los modelos $\{\mathbf{w}, \mathbf{w}_\mathrm{E}, \mathbf{w}_4\}$ para simplificar la notación.

$$\mathbf{p}_4(\,\mathbf{I_x}|\,T_4(\mathbf{I_x}) = 1, s) = \mathbf{p}_4(\,\mathbf{I_x}|\,T_4(\mathbf{I_x}) = 1) \tag{4.2}$$

Para esta tarea haremos uso de la base de datos CelebA [66], la cuál incluye anotaciones de atractivo para todas sus imágenes. Dado que algunos estudios afirman que ciertas expresiones faciales, como puede ser una sonrisa, afectan a la percepción del atractivo de una persona, introducimos un sesgo en los datos de entrenamiento con esta idea en mente. En concreto, usaremos las anotaciones binarias acerca de si la persona mostrada en la imagen está o no sonriendo, también disponibles en el dataset, para introducir un sesgo en la distribución de personas atractivas. Por lo tanto, a lo largo del experimento esta variable será nuestro atributo protegido s.

Durante el entrenamiento generamos un conjunto de datos con 36K imágenes, en el que la proporción de personas atractivas sonriendo es de un 70 %, frente a un 30 % de personas atractivas que no sonríen. A su vez, en el caso de las personas no atractivas invertimos la proporción, teniendo un 30 % de personas que sonríen frente a un 70 % que no. Balanceamos este conjunto respecto al atractivo y al género, para evitar la intromisión de estos factores en los resultados obtenidos (i.e. un 67 % de las mujeres están etiquetadas como atractivas, frente a un 27 % de los hombres). Junto a este conjunto de entrenamiento, generamos un segundo conjunto de imágenes en el que dicho sesgo no está presente, teniendo un 50 % de personas sonriendo tanto en el grupo de personas atractivas como en el que no, con el que entrenar un clasificador no sesgado que nos sirva de *baseline*.

Usando el conjunto de datos sesgado, entrenamos 3 clasificadores $\mathbf{w}_4$ distintos. Cada uno de ellos hace uso de una representación facial distinta (i.e. los *embeddings* $\mathbf{x}$ de ResNet-50 [75], así como las representaciones emocionalmente ciegas $\boldsymbol{\varphi}_\mathrm{SN}(\mathbf{x})$ y $\boldsymbol{\varphi}_\mathrm{LNTL}(\mathbf{x}')$, generadas con SensitiveNets [71] y Learning Not To Learn [53] respectivamente). En todos los casos, diseñamos el clasificador como una capa densa de 1024 unidades y activación ReLU, seguida de una capa de salida con una unidad de activación sigmoidal.

La Tabla 4.2 recoge los resultados finales de este experimento, codificando la condición de *equal of opportunity* como la diferencia entre el TPR de cada clase. Dichos resultados son la media de 5 iteraciones del experimento, generando en cada una nuevos conjuntos de entrenamiento y validación. Para evaluar cada clasificador, generamos conjuntos de validación de condiciones similares al conjunto de entrenamiento no sesgado, utilizando 4K imágenes de la partición de evaluación de CelebA. Como puede observarse, ante un conjunto de entrenamiento no sesgado, la diferencia de TPR apenas llega al 2,08 %, con una precisión del 77,26 %. En este caso, el clasificador no encuentra una correlación significativa entre que la persona sonría y su atractivo, por lo que no discrimina en base al atributo protegido, obteniendo un valor superior al 80 % en el TPR de ambas clases. Nótese que este caso obtiene tanto la mejor precisión general, como TPR del grupo de personas que no sonríen. Por contra parte, cuando entrenamos el clasificador con un conjunto sesgado y los *embeddings* $\mathbf{x}$, la diferencia de TPR asciende hasta un 17,47 %. En este caso, las representaciones de ResNet-50 extraen suficiente información

Método(dataset)	Precisión	TPR Sonr.	TPR No Sonr.	Eq. Opportunity
$\mathbf{x}$ (no sesgado)	77,26 %	84,55 %	82,47 %	2,08 %
$\mathbf{x}$ (sesgado)	76,23 %	84,17 %	66,70 %	17,47 %
$\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ (sesgado)	74,50 %	81,87 %	73,58 %	8,29 %
$\boldsymbol{\varphi}_{\text{LNTL}}(\mathbf{x}')$ (sesgado)	76,62 %	86,97 %	73,70 %	13,27 %

Tabla 4.2: Resultados de la clasificación de atractivo en CelebA [66] usando distintas representaciones faciales.

de la expresión facial para poder detectar la presencia de una sonrisa, y por lo tanto el clasificador es capaz de modelar la relación entre ambos atributos. Se puede destacar también que, al introducir el conjunto sesgado, el TPR de la clase privilegiada (i.e. $s = 1$) se ha mantenido casi invariable, explicándose el aumento de la diferencia de TPR en una caída de rendimiento en la clase no privilegiada.

Frente a estos dos casos, entrenamos el clasificador con las representaciones faciales emocionalmente ciegas que obtuvimos en la Sección 4.2. Recordemos que, entre las 6 expresiones faciales asociadas a emociones que eliminamos se encontraba la expresión "Feliz", que según el EMFACS [14] incluye la activación de la AU 12 (ver Tabla 2.1 y Figura 2.1(h)), muy cercana a lo que comúnmente se entiende por una sonrisa. Por este motivo, se puede intuir que estas representaciones serán útiles para reducir la diferencia de TPR respecto al caso entrenado con las representaciones originales de ResNet-50. Como puede observarse en la Tabla 4.2, al usar las representaciones ciegas $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ y $\boldsymbol{\varphi}_{\text{LNTL}}(\mathbf{x}')$ con el conjunto sesgado, el TPR de la clase no privilegiada pasa del 66,7 % al 73,58 % y 73,7 % respectivamente. Este aumento conlleva una reducción en ambos casos de la brecha entre el TPR de las dos clases, siendo de un 8,29 % en el caso de SN, y del 13,27 % para LNTL. Aunque la reducción de esta brecha es más notoria usando los *embeddings* de SN, es importante inspeccionar el TPR de la clase privilegiada. En el caso de SN, dicho valor se ha visto reducido al 81,87 %, casi 3 puntos menos que en el caso no sesgado. Esta reducción no es intencionada y no debería producirse, ya que para igualar el rendimiento de ambas clases lo ideal sería no penalizar a la clase privilegiada. Por su parte, en el caso de LNTL el TPR de las personas sonriendo ha aumentado hasta casi el 87 %, siendo el mayor de todos los casos.

Como conclusión, mediante el uso de representaciones faciales en las que la información referente a cierto atributo s se ha eliminado, podemos entrenar algoritmos con conjuntos de datos que incluyen un sesgo correlacionado con s reduciendo su impacto en el sistema, y su reproducción. Este diseño es bastante importante, ya que en muchas ocasiones el acceso a un dataset correctamente representativo y libre de sesgos no será posible, teniendo que recurrir a diseños más inteligentes que mitiguen los posibles efectos adversos de estas condiciones.

Capítulo 5

Conclusiones y trabajo futuro

En los últimos años, la presencia de algoritmos de toma de decisión automática ha aumentado considerablemente, jugando un papel fundamental en muchos procesos que afectan a la vida cotidiana de las personas. Ante el riesgo de que estos algoritmos vulneren la privacidad de los usuarios, o se conviertan en nuevas fuentes de discriminación, gran parte de los gobiernos del mundo han desarrollado nuevas regulaciones que protejan a los usuarios finales de los posibles efectos adversos de dichas tecnologías.

Considerando nuestro estado emocional como una información privada y potencialmente explotable por los sistemas actuales, en este trabajo nos hemos centrado en el aprendizaje de nuevas representaciones faciales en las que dicha información no esté presente, de forma que no pueda ser explotada ni por un sistema ni por terceras personas. Dado que típicamente estas representaciones son extraídas en el marco del reconocimiento facial haciendo uso de redes neuronales profundas, nuestro enfoque se ha centrado en este campo, usando de una red neuronal convolucional muy popular en los últimos años como base de nuestros experimentos.

En primer lugar, hicimos una revisión del Estado del Arte en el campo del reconocimiento de emociones a partir de imágenes faciales, analizando su evolución desde métodos tradicionales basados en características *handcrafted*, hasta los más recientes basados en tecnología *deep learning*. A su vez, repasamos los últimos avances de la comunidad científica en el desarrollo de métodos orientados a la eliminación de información sensible, y a la corrección de sesgos en los algoritmos de aprendizaje automático. Muchos de estos métodos actúan directamente en la estrategia de aprendizaje, y suelen presentar entrenamientos más complejos, como pueden ser las estrategias de entrenamiento adversarias.

Posteriormente, presentamos una formulación general del problema, y adaptamos dos métodos del Estado del Arte para generar representaciones faciales emocionalmente ciegas a partir de la información extraída por un modelo de reconocimiento facial pre-entrenado. Nuestros experimentos demostraron que la información referente a las expresiones faciales típicamente asociadas a emociones está presente en las representaciones extraídas por dicho modelo, siendo posible eliminarla aplicando los métodos propuestos, sin que esto suponga una caída significativa del rendimiento en otras tareas de análisis facial.

Finalmente, propusimos un experimento consistente en la clasificación de atractivo a partir de imágenes faciales, en el que tratamos la expresión facial como un atributo protegido que no debe influir en el proceso. Los resultados de dicho experimento sugieren que el uso de representaciones ciegas a las expresiones faciales puede ser de utilidad para mitigar el impacto de estas en la clasificación final, ayudándonos a alcanzar un criterio de equidad en presencia de datos sesgados.

Como trabajo futuro, se podría ampliar el estudio aquí presentado a otros modelos pre-entrenados de reconocimiento facial, de forma que se obtengan unos resultados más genéricos a los aquí expuestos. A su vez, dicho estudio podría incluir nuevos procedimientos, como técnicas de visualización de las activaciones intermedias, con el objeto de entender mejor la información extraída por las redes neuronales profundas. Frente a esta línea de trabajo, dada la presencia de la información referente a las expresiones faciales, y su posible efecto como fuente de variación no deseada, se podría estudiar el impacto de estas en el rendimiento final de los modelos entrenados, tanto para el reconocimiento facial como para otras tareas de análisis facial.

Glosario

IA	*Inteligencia Artificial*
CV	*Computer Vision*
NLP	*Natural Language Processing*
GDPR	*General Data Protection Regulation*
CNN	*Convolutional Neural Network*
DNN	*Deep Neural Network*
FACS	*Facial Action Coding System*
AU	*Action Unit*
EMFACS	*Emotional Facial Action Coding System*
LBP	*Local Binary Patterns*
SVM	*Support Vector Machine*
HMM	*Hidden Markov Model*
DBN	*Dynamic Bayesian Network*
NN	*Neural Network*
DBN	*Deep Belief Network*
GAN	*Generative Adversarial Network*
ReLU	*Rectified Linear Unit*
SN	*SensitiveNets*
LNTL	*Learning Not To Learn*
RF	*Random Forest*
TPR	*True Positive Rate*

Bibliografía

[1] P. Covington, J. Adams, and E. Sargin. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, New York, NY, USA, 2016.

[2] S. Chandler. The AI chatbot will hire you now. *Wired*, Sep. 2017.

[3] S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 2016.

[4] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, NY, USA, Feb 2018.

[5] B. Goodman and S. Flaxman. EU regulations on algorithmic decision-making and a "Right to explanation". *AI Magazine*, 38, Jun. 2016.

[6] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019.

[7] R. W. Picard. *Affective computing*. 2000.

[8] F. Eyben, M. Wöllmer, et al. Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. *Advances in Human-Computer Interaction*, 10 2010.

[9] E. Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[10] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

[11] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[12] P. Ekman and W. V. Friesen. *Facial action coding system: a technique for the measurement of facial movement*. Consulting Psychologist Press, Palo Alto, CA, 1978.

[13] P. Ekman, W. V. Friesen, and J. C. Hager. *The facial action coding system CD-ROM*. Reasearch Nexus, Salt Lake City, UT, 2002.

[14] P. Ekman and W. V. Friesen. EMFACS-7: Emotional Facial Action Coding System. University of California, California, 1983.

[15] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17:124–129, 1971.

[16] R. E. Jack, O. G. B. Garrod, H. Yu, R. Caldara, and P. G. Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[17] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[18] S. Argaud, M. Vérin, P. Saleau, and D. Grandjean. Facial emotion recognition in Parkinson's disease: A review and new hypotheses. *Movement disorders : official journal of the Movement Disorder Society*, 33:554–567, 2018.

[19] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing*, pages 117–124, Berlin, Heidelberg, 2013.

[20] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *ACM ICMI*, 2013.

[21] A. Dhall, O.V. Ramana Murphy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: EmotiW 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, page 423–426, New York, NY, USA, 2015.

[22] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27:803–816, 2009.

[23] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. In *IEEE transactions on pattern analysis and machine intelligence*, volume 29, pages 915–928, 2007.

[24] C. F. Benitez-Quiroz, R. Srinivasan, and A. M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[25] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111:1454–1462, 2014.

[26] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[27] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[28] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon. From individual to group-level emotion recognition: EmotiW 5.0. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, page 524–528, New York, NY, USA, 2017.

[29] A. Mollahosseini, B. Hasani, and M. H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions Affective Computing*, 10(1):18–31, January 2019.

[30] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–8, 2008.

[31] M. S. Bartlett, G. Littelwort, M. Frank, C. Lainscek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spotaneous behavior. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[32] M. Pantic and L. J. M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. In *IEEE Transactions on Systems, Man, and Cybernetics*, volume 34, pages 1449–1461, 2004.

[33] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn. Graph-preserving sparse non negative matrix factorization with application to facial expression recognition. In *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, volume 41, pages 38–52, 2011.

[34] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2562–2569, 2012.

[35] I. Cohen, N. Sebe, A. Garg, I. Chens, and T. S. Huang. Facial expression recognition from video sequences: temporal and static modeling. 91:160–187, 2003.

[36] M. Yeasin, B. Bullot, and R. Sharma. From facial expression to level of interests: a spatio-temporal approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[37] I. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1–6, 2005.

[38] Y. Tang. Deep learning using linear support vector machines. *arXiv/1306.0239*, 2013.

[39] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, page 543–550, 2013.

[40] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 2001.

[41] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, 2013.

[42] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[43] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108, 2014.

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

[45] H. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, page 443–449, 2015.

[46] B. Knyazev, Shvetsov R., N. Efremova, and A. Kuharenko. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. *arXiv/1711.04598*, 2017.

[47] H. Ding, S. K. Zhou, and R. Chellappa. FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. In *12th IEEE International Conference on Automatic Face Gesture Recognition*, pages 118–126, 2017.

[48] I. Goodfellow, J. Pouget-Abadie, et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. 2014.

[49] Y. Lai and S. Lai. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In *13th IEEE International Conference on Automatic Face Gesture Recognition*, pages 263–270, 2018.

[50] F. Zhang, T. Zhang, Q. Mao, and C. Xu. Joint pose and expression modeling for facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3359–3368, 2018.

[51] H. Yang, Z. Zhang, and L. Yin. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In *13th IEEE International Conference on Automatic Face Gesture Recognition*, pages 294–301, 2018.

[52] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

[53] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim. Learning not to learn: Training deep neural networks with biased data. In *Proc. IEEE Conf. on CVPR*, 2019.

[54] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *arXiv/2003.02488*, 2020.

[55] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proc. of EMNLP*, pages 2979–2989, Jan. 2017.

[56] M. De-Arteaga, R. Romanov, et al. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 120–128, 2019.

[57] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *arXiv/2003.02488*, 2020.

[58] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 560–568, 2008.

[59] B. Berendt and S. Preibusch. Exploring discrimination: A user-centric evaluation of discrimination-aware data mining. In *IEEE International Conference on Data Mining Workshops*, pages 344–351, 2012.

[60] T. Kehrenberg, Z. Chen, and N. Quadrianto. Tuning fairness by marginalizing latent target labels. *arXiv/1810.05598*, 2018.

[61] T. Calders and S. Verwers. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21:277–292, 2010.

[62] Y. Zhang, R. Bellamy, and K. Varshney. Joint optimization of AI fairness and utility: A human-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 400–406, 2020.

[63] N. Quadrianto, V. Sharmanska, and O. Thomas. Discovering fair representations in the data domain. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8227–8236, 2019.

[64] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3:1–3:9, 2019.

[65] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*, page 2642–2651, 2017.

[66] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2015.

[67] S. Jia, T. Lansdall-Welfare, and N. Cristianini. Right for the right reason: Training agnostic networks. In *Advances in Intelligent Data Analysis XVII*, pages 164–174, 2018.

[68] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, January 2016.

[69] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 692–702, 2019.

[70] A. Romanov, M. De-Arteaga, et al. What's in a name? reducing bias in bios without access to protected attributes. In *Proceedings of NAACL-HLT*, page 4187–4195, 2019.

[71] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana. SensitiveNets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[72] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conf. on CVPR*, pages 815–823, Jun. 2015.

[73] M. Alvi, A. Zisserman, and C. Nellaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision*, Sep. 2018.

[74] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. *arXiv/1802.06309*, 2018.

[75] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. on CVPR*, pages 770–778, Jun. 2016.

[76] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[77] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[78] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller. Learning to align from scratch. In *Advances in Neural Information Processing Systems*, 2012.

Anexo: Publicación

Learning Emotional-Blinded Face Representations

Alejandro Peña, Julian Fierrez, Aythami Morales
School of Engineering
Universidad Autonoma de Madrid, Madrid, Spain
{alejandro.penna, julian.fierrez, aythami.morales}@uam.es

Agata Lapedriza
Universitat Oberta de Catalunya, Barcelona, Spain
Massachusetts Institute of Technology, Cambridge, USA
alapedriza@uoc.edu/agata@mit.edu

***Abstract*—We propose two face representations that are blind to facial expressions associated to emotional responses. This work is in part motivated by new international regulations for personal data protection, which force data controllers to protect any kind of sensitive information involved in automatic processes. The advances in Affective Computing have contributed to improve human-machine interfaces but, at the same time, the capacity to monitorize emotional responses triggers potential risks for humans, both in terms of fairness and privacy. We propose two different methods to learn these expression-blinded facial features. We show that it is possible to eliminate information related to emotion recognition tasks, while the performance of subject verification, gender recognition, and ethnicity classification are just slightly affected. We also present an application to train fairer classifiers in a case study of attractiveness classification with respect to a protected facial expression attribute. The results demonstrate that it is possible to reduce emotional information in the face representation while retaining competitive performance in other face-based artificial intelligence tasks.**

I. Introduction

During the past 15 years there has been a lot of effort in creating technologies to extract emotional information from facial expressions [1], [2]. These facial analysis technologies can contribute to improve human-centric AI applications, like enhancing the user experience [3] or facilitating the human-computer interaction [4].

However, with the increase of image-capturing devices and available software for face image processing, face analysis technologies can also trigger potential risks for humans, both in terms of fairness and privacy. First, facial analysis software inherits human biases [5], [6], making them to perform poorly or unfairly on groups of population that are not well represented in the training data [7]. Second, humans might want to keep their emotions private or to make sure emotion recognition software is not used without their consent. Notice that privacy protection is deeply embedded in the normative framework that underlies various national and international regulations. For example, in April 2018 the European Parliament adopted a set of laws aimed to regularize the collection, storage and use of personal information [8]. In particular, these laws encourage to integrate privacy preserving methods in the technology when it is created.

As a possible solution for preserving the users privacy in the context of automatic face recognition, we propose to extract face features that are blind to facial expressions. As shown in Sec. V-A, generic face features learned for the task of subject recognition preserve information to perform tasks related to facial expression classification. However, features extracted for the target task of subject recognition do not need to preserve this facial expression information. In this paper we show that we can effectively learn alternative face feature representations for the task of subject classification that are blind to facial expression. Notice that our work is in the direction of creating automatic emotion-suppression systems, i.e., algorithms to automatically remove emotional information from captured data, with the goal of preserving privacy. A similar idea was recently explored by [9], where the goal was to suppress physiological information from facial videos. Both facial expressions and physiological signals contain information related to emotional states.

In Sec. III-A we formally describe the problem of learning the emotional-blinded face representations. Then, we propose two different methods to learn these expression-blinded facial features, which are based on existing generic techniques for learning agnostic representations. The first one (*SensitiveNets*) consists of learning a discriminator for the target task and at the same time an adversarial regularizer to reduce facial expression information. The second one (*Learning not to Learn*) consists of using a regularized loss function during learning, which quantifies the amount of information on the sensitive task (facial expression recognition) by computing the mutual information between the feature space and a pre-trained facial expression classifier. The details of these two methods can be found in Sec. III-B.

To validate the proposed framework and methods we perform an extensive set of experiments (Sec. V). First, we show that face features learned for subject verification contain significant information to perform facial expression classification (sensitive information). Then, we show that both of the proposed methods can actually eliminate information related to facial expression. In particular, for the first method, we show how the facial expression recognition accuracy drops significantly when our proposed blinded face representations are applied, while the performance of subject verification, gender recognition, and ethnicity classification are just slightly affected. Finally, our last experiment shows how the proposed methods can be applied in another face analysis problem (Attractiveness Classification) to protect the emotional information.

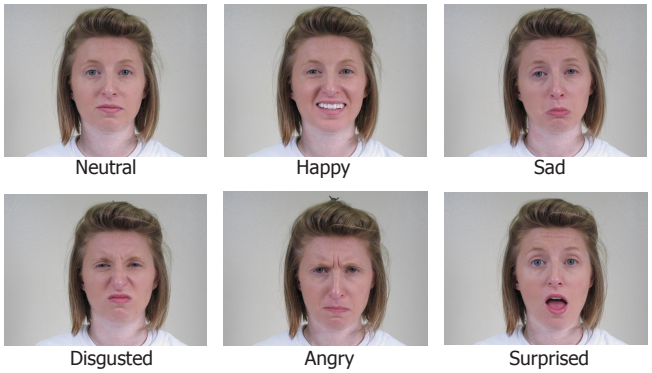


Fig. 1. Visual examples of the facial expressions, corresponding to basic emotions, that are used in our experiments. Images from CFEE database [10].

II. Related Works

The study of new learned representations to improve the fairness of learning processes has attracted other researchers [11], [12], [13], [14]. In [12], [13] researchers proposed projection methods to preserve individual information while obfuscating membership to specific groups. The main drawback of the proposed techniques was that discrimination was modelled as statistical imparity, which is not applicable when the classification goal is not directly correlated with membership in a specific group.

Bias correction and sensitive information removal are related to each other but they are not necessarily the same thing. Bias is traditionally associated with unequal representation of classes in a dataset [15]. Dataset bias can produce unwanted results in the decision-making of algorithms, e.g., different face recognition accuracy depending of your ethnicity [16], [17]. Researchers have explored new learning processes capable to compensate this dataset bias [18], [19], but the correction of biased training processes does not necessarily serve to eliminate sensitive information from the trained representation. While the correction of biased models seeks to generate representations that perform similarly for different groups or classes, the removal of sensitive information seeks to eliminate this information from that representation. The proposal in [18] is based on a joint learning and unlearning algorithm inspired in domain and task adaptation methods [20]. The authors of [21] propose a new regularization loss based on mutual information between feature embeddings and bias, training the networks using adversarial [22] and gradient reversal [23] techniques. Finally, in [24] a privacy-preserving learning method is proposed to remove sensitive information in feature embeddings, without losing performance in the main task. These works reported encouraging results showing that it is possible to remove sensitive information (named as spurious variations in [18]) for age, gender, ancestral origin, and pose in face processing for different applications [25].

On the other hand, the normalization of face images directly in the raw image space according to specific face attributes such as pose [26] or gender [27], [28] is a challenging task. In [27] researchers proposed de-identification techniques that obfuscate gender attributes while preserving face verification accuracy. The method was based on Generative Adversarial Networks trained to generate androgynous images capable of fooling gender detection systems. Similarly, the method in [26] proposed 3D models to normalize the face expressions. Although these methods showed promising results to generate realistic images, the main drawback of these techniques is that sensitive information is not eliminated but distorted. In [24], researchers demonstrated that sensitive information can be easily detected in those images when supervised learning processes are trained in the distorted domain.

A. How Emotions are Expressed in Face Images

Systems for emotion prediction from facial expression are usually based on the Facial Action Coding System [29], which encodes the facial expression using a set of specific localized face movements, called Action Units (AU). State-of-the-art systems for AU detection are based on deep learning models trained with large datasets [30]. These methods show impressive accuracies, even in uncontrolled environments [31]. However, while there are systems for AUs detection that are accurate enough to be used in practical applications, the prediction of emotions from these face movements is a more challenging problem. In that case, given a specific configuration of these face movements (that we call facial expression) the goal is to recognize the emotion category expressed by the face. There are several works that attempt to recognize the 6 basic emotions proposed by Ekman and Friesen [32] (*Anger* , *Disgust*, *Fear*, *Happiness*, *Sadness*, and *Surprise*) using AUs as a mid-level representation features [2]. These methods are based on the assumption that each basic emotion is universally expressed with a specific combinations of AUs (see Fig. 1).

On the contrary, there are studies showing that there is no universal correspondence between AUs and emotions and, therefore, it is not always possible to recognize emotions just with the information provided by facial expressions [33]. Although this lack of agreement on whether it is possible or not, in certain circumstances, to recognize emotions just from facial expressions, the studies on psychology consistently show that facial movements and expressions communicate a lot of information, including information related to emotional states [33], [34]. Thus, learning face features that are blind to facial expressions, as proposed in this paper, can actually contribute to preserve emotion privacy.

Additionally, understanding how facial expressions are represented in feature embeddings of deep neural networks models is important to gain insights into the learning processes of these algorithms. Most face recognition algorithms are trained to be agnostic to this information (i.e. facial expressions may change and these changes should not affect the recognition tasks). However, the features used to recognize a face are also useful in general to recognize face gestures. Face expression databases traditionally include both AUs and emotion labels [35]. These databases are usually employed to model face gestures as well as affective interfaces.

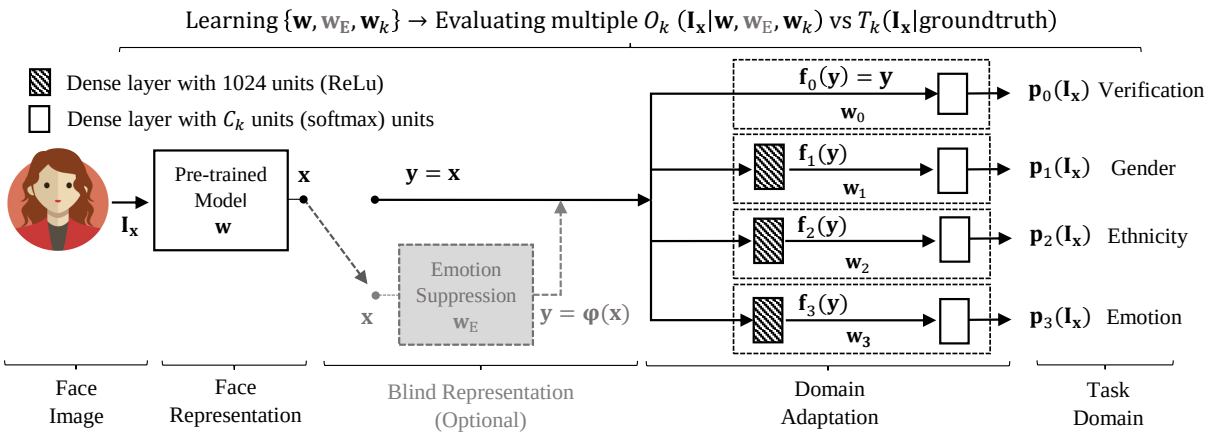


Fig. 2. General framework including domain adaptation from a pre-trained face representation to multiple tasks ($k = 0$ to 3) with and without the emotional-blinded representation $\boldsymbol{\varphi}(\mathbf{x})$. C_k is the number of classes for task k (e.g. $C_1 = 2$ corresponds to classes *male* and *female*). $\mathbf{f}_k$ is the projection to the adapted domain and $\mathbf{p}_k$ is a vector with the probabilities of each class for task k.

III. Learning Emotional-Blinded Representations

A. Problem Formulation

We employ the privacy-preserving learning framework showed in Fig. 2 and detailed in [24]. The feature vector $\mathbf{x} \in \mathbb{R}^N$ is a face representation (a.k.a face embedding) obtained as the output of one of the last layers of a face model defined by parameters $\mathbf{w} \in \mathbb{R}^M$. In our framework, the parameters of the model $\mathbf{w}$ were trained to reveal patterns associated to the identity of face images $\mathbf{I_x}$ (i.e. face verification). This pre-trained model and the databases employed for training will be detailed in Sec. V-A.

In this framework, domain adaptation is used to transform the original representation trained for face verification $\mathbf{f}_0(\mathbf{y})$ into a new representation $\mathbf{f}_k(\mathbf{y})$ ($k \geq 1$ in Fig. 2) trained for different tasks ($k = 1$: for Gender classification, $k = 2$: Ethnicity classification, $k = 3$: Emotion classification). This adaptation is performed leaving fixed $\mathbf{w} = \mathbf{w}^*$ as obtained in the pre-trained model. The domain adaptation process for a task $k \geq 1$ results in a new learned model $\mathbf{w}_k^*$ used to transform the original representation $\mathbf{x}$ for the specific task k. Given a face image, the final output of the learned model (i.e. pre-trained plus domain adaptation) is a vector $\mathbf{p}_k(\mathbf{I_x})$ containing C_k probabilities associated to each of the classes of task k.

In this work, we evaluate the face embeddings generated by the pre-trained model according to its performance in the original task (i.e. face verification) and 3 other different tasks: 1) Gender Classification; 2) Ethnicity Classification; and 3) Emotion Classification based on five of the six basic emotions proposed by Ekman plus the neutral expression (*Neutral*, *Happy*, *Sad*, *Disgusted*, *Angry*, *Surprised*).

The models $\{\mathbf{w}^*, \mathbf{w}_k^*\}$ are trained for a given task k represented by its target function T_k. The aim of the learning process is to minimize the error between the output O_k of the model and the target function T_k (e.g. $T_1 = 1$ for *male* and $T_1 = 0$ for *female*). The most popular approach for that is to train $\mathbf{w}$ and $\mathbf{w}_k$ by minimizing a loss function $\mathcal{L}_1$ over a set of Pre-training samples $\mathcal{P}$ for which we have groundtruth targets:

$$\min_{\mathbf{w},\mathbf{w}_k} \sum_{\mathbf{I_x} \in \mathcal{P}} \mathcal{L}_1[\, O_k(\, \mathbf{I_x}|\mathbf{w}, \mathbf{w}_k)\, ,\, T_k(\mathbf{I_x}|\text{groundtruth})\,] \qquad (1)$$

The parameters $\{\mathbf{w}^*,\mathbf{w}_k^*\}$, trained using Eq. (1), generate a representation $\mathbf{f}_k(\mathbf{y})$ that maximizes the performance of the model for the task k.

In this framework, the goal of emotional-blinded learning starting from pre-trained networks is to solve, including or not the Emotional Suppression module, the following problem:

$$\min_{\mathbf{w},\mathbf{w}_\text{E},\mathbf{w}_k} \sum_{\mathbf{I_x} \in \mathcal{S}} \{\mathcal{L}_1[O_k(\mathbf{I_x}|\mathbf{w}, \mathbf{w}_\text{E}, \mathbf{w}_k), T_k(\mathbf{I_x}|\text{groundtruth})] + \\ + \mathcal{L}_2[O_3(\mathbf{I_x}|\mathbf{w}, \mathbf{w}_\text{E}, \mathbf{w}_3), T_3(\mathbf{I_x}|\text{groundtruth})]\} \qquad (2)$$

where $\mathcal{L}_2$ represents a loss function intended to minimize performance in the emotion recognition task T_3 while $\mathcal{L}_1$ tries to maximize performance in a different task T_k. In our experiments we use T_0 (Face Verification) as a task to maximize the performance. In the optimization problem (2) we may use a Suppression training dataset $\mathcal{S}$ different to $\mathcal{P}$, and the optimization can take advantage of a previous solution $\{\mathbf{w}^*,\mathbf{w}_k^*\}$ to (1) in different ways. Let's denote the solution to (2) as $\{\mathbf{w}^{**},\mathbf{w}_\text{E}^{**},\mathbf{w}_k^{**}\}$.

In our experiments, we begin without Emotion Suppression ($\mathbf{y} = \mathbf{x}$ in Fig. 2) generating $\mathbf{w}^*$ in a face recognition task by pre-training using the VGGFace2 database (3 million images from more than 9,000 people [36]). We then fix $\mathbf{w}^*$ and train the Emotion classifier $\mathbf{w}_3^*$ with the CFEE database (1,380 images from 230 people, with 6 images per subject, corresponding each of these 6 images to a different emotion [10]). Finally, we solve Eq. (2) considering $\{\mathbf{w}^*,\mathbf{w}_k^*\}$ as a starting point for obtaining the solution $\{\mathbf{w}^{**},\mathbf{w}_\text{E}^{**},\mathbf{w}_k^{**}\}$ taking various optimization shortcuts as detailed in the following.

B. Suppressing Emotions from Face Representations

1) Method 1 - SensitiveNets: The work [24] recently proposed a general method to generate privacy-preserving representations starting from pre-trained networks. Here we adapt that approach to remove emotional information for the primary task k from 0 to 2 in Fig. 2.

Applying SensitiveNets to the general methodology presented before leads to: 1) fixing $\mathbf{w}^{**} = \mathbf{w}^{*}$, 2) activating the Emotion Suppression block $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ (SN for SensitiveNets) in Fig. 2, and then 3) solving the following version of Eq. (2):

$$\min_{\mathbf{w}_{\text{E}},\mathbf{w}_3} \sum_{\text{triplet}\in\mathcal{S}_{\text{P}}} \{\mathcal{L}_1[O_k(\text{triplet}|\mathbf{w}_{\text{E}},\mathbf{w}_3), T_k(\text{triplet}|_{\text{groundtruth}})] + \\ + \Delta^{\text{A}} + \Delta^{\text{P}} + \Delta^{\text{N}}\} \\ \text{s.t. } \max \text{Performance}^{k=3}_{\text{triplet}\in\mathcal{S}_{\text{E}}}(\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x}_{\text{triplet}}|\mathbf{w}_{\text{E}}),\mathbf{w}_3) \quad (3)$$

where triplet $= \{\mathbf{I}_{\text{A}}, \mathbf{I}_{\text{P}}, \mathbf{I}_{\text{N}}\}$, $\mathbf{I}_{\text{A}}$ and $\mathbf{I}_{\text{P}}$ are face images of the same person, $\mathbf{I}_{\text{N}}$ is a face image of a different person, $\mathcal{L}_1$ is the triplet loss function proposed for face recognition in [37][38], and the three Δ terms are adversarial regularizers used to measure the amount of emotion information in the learned model represented by $\mathbf{w}_{\text{E}}$:

$$\Delta = \log\{\, 1 + |0.9 - P_3(\, Neutral \,|\, \boldsymbol{\varphi}_{\text{SN}}(\mathbf{x}|\mathbf{w}_{\text{E}}), \mathbf{w}_3\,)|\,\} \quad (4)$$

The probability P_3 of observing a *Neutral* expression in the face embedding after Emotion Suppression ($\boldsymbol{\varphi}_{\text{SN}}$) is initially obtained with the pre-trained Emotion classifier $\mathbf{w}_3^{*}$, and SensitiveNets then iterates to solve Eq. (3) in order to obtain $\mathbf{w}_{\text{E}}^{**}$ (the Emotional Suppression projection) and $\mathbf{w}_3^{**}$ (an adapted Emotion classifier). In Eq. (4) $|\cdot|$ is the absolute value, and the Δ terms will tend to zero for larger P_3. Therefore, by minimizing them in Eq. (3) we force the training to output Neutral expression in general, in this way eliminating the capacity to detect expressions other than *Neutral* from the face representation $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$. In other words, we unlearn the facial features necessary to differentiate between different expressions.

On the other hand, Eq. (3) includes a constraint that will be enforced in subsequent iterations of SensitiveNets in a kind of min-max adversarial formulation [39]. Eq. (3) thus minimizes the emotion information in $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ with the Δ terms, trying to classify emotions based on $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ in the iterative learning with the optimization constraint (with decreasing success as the learning progresses), and maintaining the performance in the primary task with the tiplet loss term $\mathcal{L}_1$.

For solving Eq. (3) we apply the iterative adversarial learning approach proposed in [24] using the CFEE database [10] as $\mathcal{S}_{\text{E}}$ to retrain the emotion detector (i.e., enforcing the constraint), and the DiveFace database [24] as $\mathcal{S}_{\text{P}}$ to maintain the recognition accuracy.

The network $\mathbf{w}_{\text{E}}$ consists of three dense layers with 1024 units each layer (linear activation). After solving Eq. (3) the network $\mathbf{w}_{\text{E}}^{**}$ generates the emotional blinded representation $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$, which removes sensitive information (emotions in the present paper) while maintaining recognition performances.

2) Method 2 - Learning not to Learn: The second approach studied here to remove emotional features is based on [21]. Similar to SensitiveNets [24], this method uses a regularization algorithm to train deep neural networks, in order to prevent them from learning a known factor present in the training set irrelevant or undesired for a given primary task. Here we propose to unlearn emotional features for the primary task k from 0 to 2 in Fig. 2.

In this case the Emotion Suppression switch is off, therefore there is no $\mathbf{w}_{\text{E}}$, and we start from pre-trained $\{\mathbf{w}^{*}, \mathbf{w}_k^{*}, \mathbf{w}_3^{*}\}$.

The training algorithm uses a regularization loss that includes the mutual information between emotions and feature embeddings $\mathbf{x}$. These embeddings are then fed into both the main classification task network (corresponding to k from 0 to 2), and the emotion classification network $\mathbf{p}_3$. The function to optimize for emotion removal is then:

$$\min_{\mathbf{w},\mathbf{w}_k} \sum_{\mathbf{I}_{\mathbf{x}}\in\mathcal{S}} \{\mathcal{L}_c[\, O_k(\mathbf{I}_{\mathbf{x}}|\mathbf{w},\mathbf{w}_k)\,,\, T_k(\mathbf{I}_{\mathbf{x}}|_{\text{groundtruth}})\,] + \\ + \lambda\mathcal{I}[\,\mathbf{p}_3(\mathbf{I}_{\mathbf{x}})\,;\,\mathbf{x}\,]\,\} \quad (5)$$

where $\mathcal{L}_c$ denotes the cross-entropy loss, $\mathcal{I}$ represents the mutual information and λ is an hyper-parameter.

To compute the mutual information in Eq. (5), we used the emotion classification network to approximate the a posteriori distribution of the emotional classifier $\mathbf{p}_3(\mathbf{I}_{\mathbf{x}})$. The training algorithm can be implemented in practice following an adversarial strategy [22], combined with the use of the gradient reversal technique [23].

IV. Data and experimental set up

To obtain the face representation $\mathbf{x}$ we use a learning architecture with state-of-the-art performance in face recognition tasks: ResNet50, proposed in [40]. ResNet50 has around 41M parameters split in 34 residual layers. The pre-trained model used in this work was trained from scratch with VGGface2 dataset [36]. This ResNet50 model achieved 98.0% accuracy in face verification with the IJB-A dataset [41].

Using the base representation $\mathbf{x}$ generated by the pre-trained network ResNet50 we trained different classifiers as depicted in Fig. 2 according to the following labeled databases:

- DiveFace [24]: The DiveFace database contains annotations equitably distributed among 6 demographic classes, related to gender and 3 ethnic groups (East Asian | *Sub-Saharan* and *South Indian* | *Caucasian*), with 24K different identities and a minimum of 3 images per identity. This database was used to train the emotional-blinded representation. Additionally, 12K subjects of this database were used to train and test the gender and ethnicity classification.
- CFEE [10]: The Compound Facial Expressions of Emotion database includes facial images of 230 different users. For every user, we selected an image belonging to

each of the 22 categories present in the dataset: 6 basic emotions, 15 compound emotions (i.e. a combination of two basic emotion), and neutral expression. All images represent a fully recognizable expression, being captured in a controlled environment of illumination and pose. We used the 6 basic emotion of this database to train the emotional-blinded representation.

- LFW [42]: Labeled Faces in the Wild is a database for research on unconstrained face recognition. It contains more than 13K images of faces collected from the web. We employ the aligned images [43] from the test set provided with view 1 and its associated evaluation protocol.
- CelebA [44]: The CelebA dataset has a total of 202K celebrity images from more than 10K identities. Each image is annotated with 40 binary attributes, including appearance features, gender, age, attractiveness and emotional state, and 5 landmark positions. The dataset is partitioned into 2 splits, with 8K identities retained as the training set, and the remaining 2K as the test set.

In order to measure how much emotional information is available in the face representation, we trained different emotion classifiers using either original embeddings $\mathbf{x}$ or emotional-blinded representations $\boldsymbol{\varphi}(\mathbf{x})$. We measured the amount of emotional information as the performance achieved by these classification algorithms. We assume that emotional information is removed by our blinding transformation $\boldsymbol{\varphi}(\cdot)$ when a significantly drop of performance in emotion classification occurs in comparison to the original emotion classification accuracy before applying that transformation.

The face recognition accuracy is obtained according to the evaluation protocol of the popular benchmark of LFW [42]. For the rest of tasks, we used 80% of the samples for training and 20% for testing. Implementation details: 150 epochs, Adam optimizer (learning rate $= 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$), and batch size of 128 samples.

V. Experiments

A. Are Facial Expressions Encoded in Generic Face Representations?

In order to better understand how the emotional features are embedded in the deep face representations, in this section we explore how identity and emotional information are represented in $\mathbf{x}$ and $\mathbf{f}_3(\mathbf{x})$.

Fig. 3 shows the two-dimensional t-SNE projection of the original face representation $\mathbf{x}$ and the learned representation $\mathbf{f}_3(\mathbf{x})$ for emotion recognition using the CFEE database [10] (detailed in Sec. IV). This database is interesting for this study because of its controlled acquisition environment (covariates such as pose or illumination are not present) and the multiple face gestures available for 230 subjects. We ran t-SNE over $\mathbf{x}$ and $\mathbf{f}_3(\mathbf{x})$ without using the emotion labels available, and then show in Fig. 3 the resulting t-SNE projections with emotion labels *a posteriori* for visualization purposes.

As we can see in Fig. 3 (Up), the projection in the original representation ignores the emotional features. The representation learned for face verification deprecates emotional features

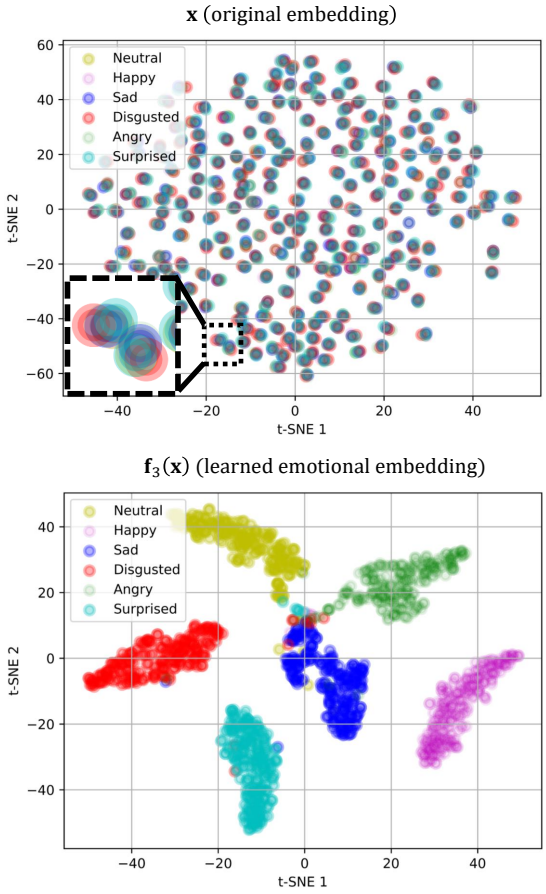


Fig. 3. t-SNE plot of the original embedding $\mathbf{x}$ (Up) and emotion feature transformation $\mathbf{f}_3(\mathbf{x})$ (Down) of the face images from CFEE database.

in order to maximize accuracy in face recognition. Face expressions can be seen as distortions that should be removed from the decision-making of the representation. However, if we freeze the weights of the ResNet model that produced the representation $\mathbf{x}$ and we train the representation $\mathbf{f}_3(\mathbf{x})$ for Emotion Classification, we can observe in Fig. 3 (Down) how emotional features were available in $\mathbf{x}$ and a simple training procedure with hundreds of samples allows to extract that information and correctly classify the emotions for more than 90% of the face images. Note that as mentioned before, ResNet was trained originally for identity recognition and these emotional features were not intentionally included in the learning process. These results illustrate that emotional information is embedded in $\mathbf{x}$ even though that representation was trained for a different purpose (i.e. face verification).

To gain insight into how the emotional features are embedded in the original representation $\mathbf{x}$, we have evaluated the performance of an emotion classifier when different amount of features from $\mathbf{x}$ are available to train $\mathbf{f}_3(\mathbf{x})$. To do this, in each iteration we randomly suppress a percentage of features of the representation $\mathbf{x}$ and we re-train the emotion representation $\mathbf{f}_3(\mathbf{x})$, always freezing the ResNet model. Fig. 4 shows the performance decay for Emotion Classification related to the

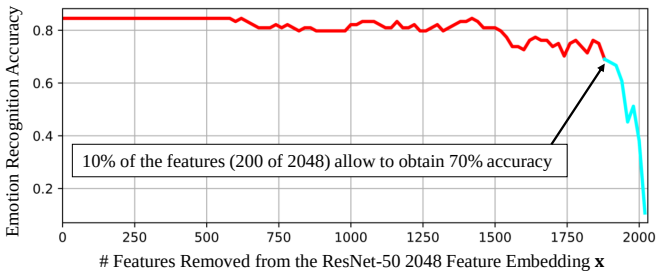


Fig. 4. Performance of the emotion classifier $\mathbf{p}_3(\mathbf{I}_\mathbf{x})$ vs number of features from $\mathbf{x}$ used to train $\mathbf{w}_3$ with fixed $\mathbf{w}$ (see Fig. 2).

TABLE I
ACCURACY OF DIFFERENT CLASSIFIERS TRAINED WITH $\mathbf{x}$ (BEFORE) OR $\boldsymbol{\varphi}(\mathbf{x})$ (AFTER). DIFF IS THE ACCURACY DROP RELATIVE TO RANDOM CHOICE (DIFF=100% REPRESENTS A RANDOM CHOICE CLASSIFIER): DIFF = (BEFORE − AFTER)/(BEFORE − RANDOM CHOICE)

Information Domain	$\mathbf{x}$	$\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$	Diff. SN	$\boldsymbol{\varphi}_{\text{LnL}}(\mathbf{x})$	Diff. LnL
ID	96.8	96.3	↓ 1%	59.4	↓ 75.0%
Gender	99.2	98.9	↓ 1%	72.7	↓ 53.9%
Ethnicity	98.8	98.6	↓ 1%	67.4	↓ 47.9%
Emotion (NN)	88.1	59.6	↓ 40%	41.6	↓ 65.0%
Emotion (SVM)	88.1	16.7	↓ 100%	25.0	↓ 88.2%
Emotion (RF)	77.4	58.3	↓ 31%	44.7	↓ 53.8%

number of features suppressed from the original representation $\mathbf{x}$. It is remarkable how well the emotion representation is capable of classifying with 70% accuracy even if the number of features available is only 10% of the original size. The model is able to keep almost the same performance until 90% of features are suppressed. This demonstrates that emotional features are latent in almost all features of the original representation $\mathbf{x}$.

B. Emotional-Blinded Face Representations

The aim is to maintain the recognition capability in other tasks related to face analysis while removing the emotion information embedded in the face representation $\mathbf{x}$ using the methods described in Sec. III-B for generating $\boldsymbol{\varphi}(\mathbf{x})$ (see Fig. 2). To analyze the effectiveness of those blinding methods, we conducted experiments on 3 datasets (see Sec. IV): DiveFace, CFEE, and LFW.

a) *Objective* 1 - *Maintaining face representation information:* the goal is to maintain the performance of the emotional-blinded face representation for other tasks different to emotion classification. We calculated the performance of 3 different face-based machine learning tasks using either original embeddings $\mathbf{x}$ or their projections $\boldsymbol{\varphi}(\mathbf{x})$. The tasks are evaluated according to the classification accuracy obtained in the test set. Table I shows the classification accuracy of representations generated by the pre-trained model before and after the projections $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ obtained by the *Method 1* and $\boldsymbol{\varphi}_{\text{LnL}}(\mathbf{x})$ obtained using the *Method 2* (see Sec. III-B). The results of the projection $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ show a very small drop of performance when the projection is applied in the first domains (ID, Gender, and Ethnicity), which demonstrates the success of our method in preserving most of the discriminative information in the face representation. The drop of performance in the method based on the $\boldsymbol{\varphi}_{\text{LnL}}(\mathbf{x})$ projection is higher for the primary tasks (ID, Gender, Ethnicity) and the emotion classification. This decay may be caused because of the disentanglement of primary and secondary tasks managed by the mutual information regularizer in Eq. (5). The method proposed in [21] was originally evaluated for problems with limited number of classes and face recognition requires feature spaces capable of allocating large number of classes (one per identity).

b) *Objective* 2 - *Removing emotional information:* to analyze the amount of emotional information available in the face representations we train different emotion classification algorithms (NN = Neural Networks, SVM = Support Vector Machines, and RF = Random Forests) either on original embeddings $\mathbf{x}$ or on their projections $\boldsymbol{\varphi}(\mathbf{x})$. Table I shows the accuracies obtained by each classification algorithm before and after the projections. Results show a quite significant drop of performance in classification when both blinding representations are applied, which demonstrates the success in reducing the emotion information from the embeddings. The emotional features are deeply embedded in the representations and in order to maintain the performance of other tasks (like in the first 3 rows of Table I), not all the emotion information was removed.

There are differences between the performances obtained by the two blinding methods. While $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ maintains higher performance in the primary tasks (ID, Gender, and Ethnicity), the emotion suppression is higher in $\boldsymbol{\varphi}_{\text{LnL}}(\mathbf{x})$. This higher suppression obtained by $\boldsymbol{\varphi}_{\text{LnL}}(\mathbf{x})$ may be due to the weaker representations generated by this method which lead to worse performance in the primary tasks. However, the accuracy obtained for emotion classification using both methods (lower than 60% in all cases) may be low enough to prevent its unwanted exploitation. Emotion-related privacy is not fully granted, but clearly improved.

Fig. 5 shows the two-dimensional t-SNE projection similar to Fig. 3 (Down) but for the emotional blinded representation $\mathbf{f}_3(\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x}))$. The results show how the domain adaptation training of $\mathbf{w}_3$ (see Fig. 2) was not able to find a representation capable of discriminating emotions in the learned representation $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$.

C. Blind Representations: Towards Equality of Opportunity

Inspired in the experiments performed in [7] for analyzing biases and achieving a specific fairness criterion, here we study how blind representations can improve the *Equality of Opportunity* [45]. For this purpose we introduce task $k = 4$: binary Attractiveness classification (*Attractive* | *Not Attractive*) based on an input face image $\mathbf{I}_\mathbf{x}$.

In this experiment, the outcome of an Attractiveness classifier with input $\mathbf{x}$ and parameters $\mathbf{w}_4$ given its positive class should be independent to the feature s we want to protect in

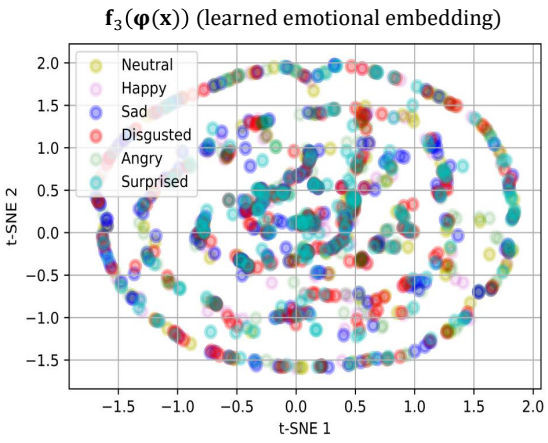


Fig. 5. t-SNE plot of the emotional embedding $\mathbf{f}_3$ trained with the proposed emotional-blinded representation $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ as input (see Fig. 2) over the CFEE database. See previous Fig. 3 (Down) for comparison.

terms of fairness. In our experiments, the protected attribute is an specific face gesture: smile. Therefore, in our case: $s \in \{Smiling, Not\ Smiling\}$.

Using the framework presented in Sec. III-A summarized in Fig. 2, the Equality of Opportunity results in: $\mathbf{p}_4(\mathbf{I_x}|\mathbf{w}^*, \mathbf{w}_4^*, T = 1, s) = \mathbf{p}_4(\mathbf{I_x}|\mathbf{w}^*, \mathbf{w}_4^*, T = 1)$. This criterion implies equal True Positive Rates across the different face gestures defined by s and the Attractiveness classifier defined by the parameters $\mathbf{w}^*, \mathbf{w}_4^*$.

We used 40K images from CelebA dataset [44], previously introduced in Sec. IV, to train the Attractiveness classifier. Some studies suggest that face expressions such as smile affect the perception of attractiveness, so we train a classifier introducing an emotional bias with this fact in mind. In our experiment, we employed the smiling annotation available in CelebA as a face gesture commonly associated to a positive emotion that can therefore introduce undesired bias. We generated an emotionally biased training set where the proportion of attractive people smiling and not smiling was 70% and 30% respectively. We introduced the opposite bias for the unattractive group with 30% and 70% of smiling and not smiling respectively. In order to avoid the appearance of other biases, we balanced the dataset in terms of attractiveness and gender, compensating the gender bias of the dataset (i.e. the proportion of attractive females is 67%, while for males is 27%). We also generated an unbiased dataset with 50% smiling and not smiling samples (randomly chosen and balanced with respect to gender).

The results in Table II show higher True Positive Rates (TPR) for the privileged class (*Smiling* in our experiment) in comparison with the non-privileged class (*Not Smiling*). The face gesture *Smiling* was irrelevant to classify the attractiveness (i.e. there was no correlation between the attributes *Smiling* and *Attractiveness*). However, a classifier trained on face embeddings $\mathbf{x}$ generated by pre-trained models like ResNet50, tends to reproduce the bias introduced in the training datasets.

TABLE II
RESULTS ON ATTRACTIVENESS CLASSIFICATION (ACC = ACCURACY). EQUAL OPPORTUNITIES IS CALCULATED AS: 100 − (TPR SMILING − TPR NOT SMILING). TPR = TRUE POSITIVE RATE IN ATTRACTIVENESS CLASSIFICATION

Method (training)	Acc.	TPR Smil.	TPR Not Smil.	Eq. Opp.
$\mathbf{x}$ (unbiased)	77.26%	84.55%	82.47%	97.93%
$\mathbf{x}$ (biased)	76.23%	84.17%	66.70%	82.53%
$\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ (biased)	74.50%	81.87%	73.58%	91.71%
$\boldsymbol{\varphi}_{\text{LnL}}(\mathbf{x})$ (biased)	76.62%	86.97%	73.70%	86.73%

Table II shows how the blind representations $\boldsymbol{\varphi}_{\text{SN}}(\mathbf{x})$ and $\boldsymbol{\varphi}_{\text{LnL}}(\mathbf{x})$ presented in Sec. III-B significantly reduce the gap between both classes by improving equality in 9% and 4% respectively. The blind representations avoid the network to exploit the latent variable related with the face gesture and reduce the impact of the biased training dataset.

Implementation details: the classifiers were composed by one fully connected layer (1024 units and ReLu activation) and one output unit (sigmoid activation), which we feed with face embeddings generated with the methods mentioned above. We repeated the experiment five times, using different training sets with 36K images from CelebA, and evaluating the resulting classifiers on validation sets with 4K images, selected from the CelebA's evaluation split.

VI. CONCLUSIONS

The growth of emotion recognition technologies has allowed great advances in fields related to human-machine interaction. At the same time, automatic systems capable to read emotions without explicit consent trigger potential risks for humans, both in terms of fairness and privacy. In this work we have proposed two face representations that are blind to facial expressions associated to emotional responses.

In addition to a general formulation of the problem, we have adapted two existing methods for this purpose of generating emotional-blinded face representations: SensitiveNets [24] and Learning not to Learn [21]. The results show that it is possible to reduce dramatically the performance of emotion classifiers (more than 40%) while the performance in other face analysis tasks (verification, gender, and ethnicity recognition) is only slightly reduced (less than 2%).

Finally, we included an experiment on facial attractiveness classification to show how to treat facial expression as protected information in face classification problems. The results show how blinded representations can improve a specific fairness criterion based on the principles and methods studied in the present paper.

VII. ACKNOWLEDGMENTS

This work has been supported by projects: PRIMA (H2020-MSCA-ITN-2019-860315), TRESPASS-ETN (H2020-MSCA-ITN-2019-860813), IDEA-FAST (IMI2-2018-15-853981), BIBECA (RTI2018-101248-B-I00 MINECO/FEDER), REAVIPERO (RED2018-102511-T), RTI2018-095232-B-C22

MINECO, and Accenture. A. Peña is supported by a research fellowship (PEJ2018-004094A) from the Spanish MINECO.

REFERENCES

[1] M. Pantic and L. J. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, 2000.

[2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 2010, 2020.

[3] F. Eyben, M. Wöllmer *et al.*, "Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car," *Advances in Human-Computer Interaction*, 10 2010.

[4] S. S. Guillén, L. L. Iacono, and C. Meder, "Affective robots: Evaluation of automatic emotion recognition approaches on a humanoid robot towards emotionally intelligent machines," *Energy*, vol. 3643, 2018.

[5] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[6] G. Bijlstra, R. W. Holland, R. Dotsch, K. Hugenberg, and D. H. Wigboldus, "Stereotype associations and emotion recognition," *Personality and Social Psychology Bulletin*, vol. 40, no. 5, pp. 567–577, 2014.

[7] N. Quadrianto, V. Sharmanska, and O. Thomas, "Discovering fair representations in the data domain," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[8] EU 2016/679 (General Data Protection Regulation). European Commission. [Online]. Available: https://gdpr-info.eu/

[9] W. Chen and R. W. Picard, "Eliminating physiological information from facial videos," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2017, pp. 48–55.

[10] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[11] S. Jia, T. Lansdall-Welfare, and N. Cristianini, "Right for the right reason: Training agnostic networks," in *Advances in Intelligent Data Analysis XVII*, W. Duivesteijn, A. Siebes, and A. Ukkonen, Eds., 2018, pp. 164–174.

[12] E. Raff and J. Sylvester, "Gradient reversal against discrimination: A fair neural network learning approach," in *IEEE International Conference on Data Science and Advanced Analytics*, 2018, pp. 189–198.

[13] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, p. 325–333.

[14] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan, "Algorithmic discrimination: Formulation and exploration in deep learning-based face biometrics," in *Proc. of AAAI Workshop on SafeAI*, Feb. 2020.

[15] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1521–1528.

[16] B. F. Klare *et al.*, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[17] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic Bias in Biometrics: A Survey on an Emerging Challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89—-103, 2020.

[18] M. Alvi, A. Zisserman, and C. Nellaaker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *European Conference on Computer Vision Workshops*, 2018.

[19] S. Nagpal, M. Singh, R. Singh, M. Vatsa, and N. Ratha, "Deep Learning for Face Recognition: Pride or Prejudiced?" *arXiv:1904.01219*, 2019.

[20] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *The IEEE International Conference on Computer Vision*, December 2015.

[21] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2672–2680.

[23] Y. Ganin, E. Ustinova *et al.*, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, p. 2096–2030, 2016.

[24] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "SensitiveNets: Learning agnostic representations with application to face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[25] A. Peña, I. Serna, A. Morales, and J. Fierrez, "Bias in multimodal AI: Testbed for fair automatic recruitment," in *IEEE CVPR Workshop on Fair, Data Efficient and Trusted Computer Vision*, 2020.

[26] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[27] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images," in *International Conference on Biometrics*, 2018, pp. 82–89.

[28] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial soft biometrics for recognition in the wild: Recent works, annotation and COTS evaluation," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 8, pp. 2001–2014, August 2018.

[29] P. Ekman and W. V. Friesen, *Facial action coding system: a technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press, 1978.

[30] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *IEEE International Conference on Computer Vision & Pattern Recognition*, 2016.

[31] C. F. Benitez-Quiroz, Y. Wang, and A. M. Martinez, "Recognition of action units in the wild with deep nets and a new global-local loss," in *International Conference on Computer Vision*, 2017, pp. 3990–3999.

[32] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124, 1971.

[33] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[34] J. A. DeVito, S. O'Rourke, and L. O'Neill, *Human Communication*. Longman, 2000.

[35] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[36] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *International Conference on Automatic Face and Gesture Recognition*, 2018.

[37] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[38] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[39] J. Wang, T. Zhang *et al.*, "Towards a unified min-max framework for adversarial exploration and robustness," *arXiv:1906.03563*, 2019.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[41] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1931–1939.

[42] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[43] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller, "Learning to align from scratch," in *Advances in Neural Information Processing Systems*, 2012.

[44] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *International Conference on Computer Vision*, 2015.

[45] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems 29*, 2016.