INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 3, No 2 (2022)
ISSN 2695-6578

Sobre el reconocimiento de emociones y la precisión de los clasificadores

About emotion recognition and the accuracy of classifiers

Maria Francesca Roig-Maimó
Departamento de Ciencias Matemáticas e Informática
Universitat de les Illes Balears
Palma, España
xisca.roig@uib.es

Miquel Mascaró-Oliver
Departamento de Ciencias Matemáticas e Informática
Universitat de les Illes Balears
Palma, España
miquel.mascaro@uib.es

Esperança Amengual-Alcover
Departamento de Ciencias Matemáticas e Informática
Universitat de les Illes Balears
Palma, España
eamengual@uib.es

Ramon Mas-Sansó
Departamento de Ciencias Matemáticas e Informática
Universitat de les Illes Balears
Palma, España
ramon.mas@uib.es

Recibido: 03.10.2022 | Aceptado: 29.11.2022

Palabras Clave

Reconocimiento de expresiones faciales
Reconocimiento de emociones
Aprendizaje automático
Dataset de expresiones faciales
Redes neuronales convolucionales
LIME

Resumen

El reconocimiento de expresiones faciales es un tópico muy popular desde el surgimiento de la Inteligencia Artificial. Desde el punto de vista de la Interacción Persona-Ordenador también existe un gran interés, ya que la expresión facial detectada aporta una gran cantidad de información a la hora de discernir la emoción que transmite. Hoy en día, las redes neuronales son uno de los sistemas de aprendizaje computacional más utilizados para reconocer y analizar emociones y, en general, los principales esfuerzos se dirigen a entrenar modelos a partir de expresiones faciales que alcancen el máximo rendimiento posible en términos de precisión. Pero, en realidad, los humanos no son perfectos en el reconocimiento de emociones a partir de expresiones faciales estáticas. En este trabajo, planteamos la cuestión de si el objetivo a perseguir al entrenar tales modelos debe basarse únicamente en alcanzar altos valores de precisión de reconocimiento o si también debemos centrarnos en tratar de emular el comportamiento humano. Intentamos llegar a una respuesta a través de comparar los resultados de clasificación de emociones en dos experimentos: uno con participantes humanos y otro con una red neuronal convolucional.

Keywords

Facial expression recognition
Emotion recognition
Machine learning
Facial expression dataset
Convolutional neural networks
LIME

Abstract

The recognition of facial expressions is a very popular topic since the emergence of Artificial Intelligence. From the point of view of Human-Computer Interaction, there is also great interest since the detected facial expression provides a large amount of information to recognize its associated emotion. Nowadays, neural networks are one of the most widely used computational learning systems for recognizing and analyzing emotions from facial expressions. Usually, the main efforts are directed towards obtaining the maximum performance of the system in terms of high accuracy. But in reality, humans are not that good at distinguishing between emotions from static facial expressions. In this paper, we raise the question of whether the goal to pursue when training models should be based solely on achieving high values of accuracy or whether we should also focus on trying to emulate human behavior. We try to provide an answer through two experiments: one with human participants and the other one with a convolutional neural network.

1. Introducción

El reconocimiento de expresiones faciales es un tópico sobre el que se ha escrito mucho desde el surgimiento de la Inteligencia Artificial (IA) (Bettadapura, 2012). Desde el punto de vista de la Interacción Persona-Ordenador (IPO) también existe un creciente interés, ya que, junto con el análisis de posturas y gestos, las emociones transmitidas a través de las expresiones faciales aportan una gran cantidad de información a la hora de tratar la comunicación no verbal (Cowie et al., 2001).
Existen aplicaciones de IPO en las que sería deseable tener en cuenta el estado emocional del usuario. Tal sería el caso, por ejemplo, de las aplicaciones que cumplen una función social: como las aplicaciones de tutoría automática, las aplicaciones de aprendizaje, los asistentes personales o las aplicaciones para el cuidado de personas mayores, donde la capacidad de reconocer las emociones del usuario podría mejorar su usabilidad.

Hoy en día, las redes neuronales son uno de los sistemas de aprendizaje automático más utilizados para reconocer y analizar la expresión humana. Sin embargo, el entrenamiento de dichos sistemas requiere de grandes *datasets* que deben estar correctamente etiquetados para garantizar altas tasas de precisión. Afortunadamente, hay una cantidad considerable de *datasets* disponibles que se pueden usar para entrenar y evaluar el rendimiento de los nuevos modelos de reconocimiento (Lucey et al., 2010; Lundqvist, Flykt, & Öhman, 1998; Lyons, Kamachi, & Gyoba, 1997; Mollahosseini, Hasani, & Mahoor, 2017; Valstar, Pantic, & others, 2010). Recientemente, el uso de *datasets* sintéticos se está popularizando, ya que proporcionan grandes cantidades de imágenes generadas sintéticamente, etiquetadas automáticamente y libres de problemas de privacidad de datos (Colbois, de Freitas Pereira, & Marcel, 2021; Oliver & Amengual Alcover, 2020). Las imágenes con un aspecto muy realista, como las que se generan con las redes generativas antagónicas (GAN) y las secuencias dinámicas con aspecto de avatar utilizadas en las aplicaciones de realidad virtual, han demostrado ser una buena sustitución de las imágenes reales, ya que obtienen tasas de reconocimiento similares a las obtenidas con *datasets* de imágenes reales (Colbois et al., 2021; del Aguila et al., 2021).

Las tasas de reconocimiento de emociones de los modelos de aprendizaje automático cuando se usan imágenes reales oscilan entre el 70% y el 95% (Ramis, Buades, Perales, & Manresa-Yee, 2022) y cuando se usan imágenes sintéticas estas tasas pueden incluso superarse (hasta un 99%) (Chirra, Uyyala, & Kolli, 2021). Habitualmente, los esfuerzos en la creación de nuevos modelos de aprendizaje automático están relacionadas con la mejora de los índices de reconocimiento de los modelos ya existentes, normalmente expresados en términos de precisión, con el objetivo de conseguir entrenar un modelo que sea capaz de realizar una clasificación perfecta. Pero ¿son estas altas tasas de precisión perseguidas por los modelos de aprendizaje automático equiparables a cómo los humanos realmente reconocen las emociones cuando miran la imagen de una expresión facial estática sin contexto?

La primera propuesta para determinar si un sistema se podía considerar inteligente data del 1950 y es conocido como el *Test de Turing*. Este test fue presentado por Alan Turing en 1950 (Turing, 1950) y consiste en determinar si una máquina puede demostrar la misma inteligencia que un ser humano en términos de generar los mismos resultados que generaría un humano. Este mismo principio de definir la inteligencia artificial como sistemas que actúan como humanos fue también establecido por Russell & Norvig en 1995 en su definición universal de lo que es la Inteligencia Artificial (Russell & Norvig, 1995). Según estas premisas, podría considerarse que el objetivo original de los algoritmos de aprendizaje automático es crear una inteligencia artificial que se comporte como los humanos. Por este motivo, en este artículo, además de centrarnos únicamente en las medidas de rendimiento para validar un modelo, nos preguntamos si un análisis de la similitud entre el comportamiento de reconocimiento del modelo y el comportamiento de reconocimiento humano también podría ser una buena medida a tener en cuenta para validar un modelo.

Con este propósito en mente, utilizamos la base de datos de expresiones faciales sintéticas UIBVFED (Oliver & Amengual Alcover, 2020) para comparar el rendimiento obtenido por un algoritmo de aprendizaje automático y el rendimiento obtenido por participantes humanos al reconocer expresiones faciales sintéticas estáticas. Realizamos dos experimentos: el primero con participantes humanos y el segundo con una red neuronal convolucional.

Los resultados obtenidos muestran que, contrariamente al objetivo que se persigue cuando se entrena un modelo de aprendizaje automático, los humanos no son muy buenos para distinguir entre diferentes emociones cuándo la única información disponible está formada por imágenes de expresiones faciales estáticas y fuera de contexto; incluso un modelo simple de red neuronal obtiene mejores resultados. Por eso, planteamos si el objetivo al entrenar un modelo de aprendizaje automático debería ser conseguir una clasificación perfecta en términos de precisión o bien conseguir emular el proceso de clasificación humano, incluyendo también sus confusiones. ¿Deberíamos conseguir un clasificador de expresiones faciales infalible o deberíamos conseguir un clasificador de expresiones faciales que fuera indistinguible de un humano, es decir, que fuera capaz de superar el *Test de Turing*?

El resto de este artículo está organizado de la siguiente manera: la Sección 2 presenta el *dataset* utilizado para realizar los experimentos. La Sección 3 describe el experimento realizado con participantes humanos para analizar como identifican las emociones a partir de expresiones faciales estáticas de avatares sintéticos. La Sección 4 describe el experimento realizado con una red neuronal convolucional. La Sección 5 discute los resultados obtenidos en los dos experimentos y, finalmente, la última sección resume las conclusiones del trabajo.

2. El *dataset* UIBVFED

Para comprender y comparar cómo se comportan las máquinas y los humanos al reconocer e identificar emociones a partir de expresiones faciales sintéticas estáticas, utilizamos la base de datos UIBVFED (Oliver & Amengual Alcover, 2020). UIBVFED es la primera base de datos compuesta enteramente por avatares sintéticos que categoriza hasta 32 expresiones faciales. El *dataset* se compone de 640 imágenes faciales de 20 personajes virtuales. Los avatares representan a 10 hombres y 10 mujeres de diferentes etnias, de entre 20 y 80 años. Cada uno de los avatares realiza 32 expresiones faciales que se clasifican en base a las seis emociones universales (ira, asco, miedo, alegría, tristeza y sorpresa) más la emoción neutral. La Figura 1 muestra la expresión facial de sonrisa desleal (asociada a la emoción de alegría) de algunos de los personajes de la base de datos.

Figura 1: La expresión facial de la sonrisa desleal.

Las expresiones de los avatares se han generado utilizando deformaciones que siguen estrictamente las especificaciones presentes en la literatura para las unidades de acción (*Action Units*) que describen la configuración del sistema de código de acción facial (FACS) (Ekman & Friesen, 1978). De esta forma, el etiquetado de las emociones es objetivo y se obtiene directamente de tales configuraciones. En la Tabla 1 podemos observar la equivalencia entre una expresión facial y su emoción asociada.

3. El experimento con humanos

En esta sección detallamos el experimento realizado con participantes humanos para analizar cómo los humanos reconocen las emociones al observar una imagen de un avatar realizando una expresión facial.

3.1 Participantes

Veintidós participantes (3 mujeres) fueron reclutados del campus de la universidad local.

Tabla 1: Expresiones faciales asociadas a emociones

Emoción	Expresión facial
Neutral (*neutral*)	Neutral (*neutral*)
Ira (*anger*)	Ira con labios comprimidos (*enraged compressed lips*), ira gritando (*enraged shouting*), enojo (*mad*), ira severa (*anger*)
Asco (*disgust*)	Desdén (*disdain*), asco (*disgust*), repulsión (*physical repulsion*)
Miedo (*fear*)	Miedo (*afraid*), terror (*terror*), muy asustado (*very frightened*), preocupado (worried)
Alegría (*joy*)	Risa falsa 1 (*false laughter 1*), sonrisa falsa (*false smile*), sonrisa con la boca cerrada (*smiling closed-mouth*), sonrisa con la boca abierta (*smiling open-mouthed*), sonrisa sofocada (*stifled smile*), risa (*laughter*), risa estruendosa (*uproarious laughter*), risa falsa 2 (*false laughter 2*), sonrisa avergonzada (*abashed smile*), sonrisa ansiosa (*eager smile*), sonrisa congraciadora (*ingratiating smile*), sonrisa astuta (*sly smile*), sonrisa melancólica (*melancholy smile*), sonrisa desleal (*debauched smile*)
Tristeza (*sadness*)	Llorando con la boca cerrada (*crying closed mouth*), llorando con la boca abierta (*crying open-mouthed*), miserable (*miserable*), casi llorando (*nearly crying*), triste (*sad*), tristeza reprimida (*supressed sadness*)
Sorpresa (surprise)	Sorpresa (*surprise*)

Las edades oscilaron entre 24 y 53 años con una media de 28.64 años (*SD* = 6.54). No hubo requisitos de experiencia previa para participar en el estudio.

3.2 El cuestionario

El cuestionario estaba formado por cuarenta y dos preguntas. Cada pregunta consistía en una imagen de un avatar sintético del *dataset* UIBVFED realizando una expresión facial y contenía siete opciones de respuesta correspondientes a las 6 emociones universales más la neutral: ira, asco, miedo, alegría, tristeza, sorpresa y neutral (ver Figura 2).

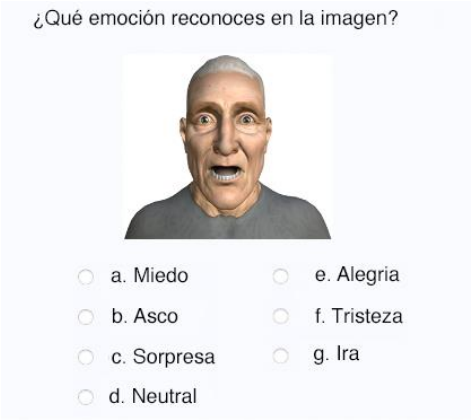

Figura 2: Ejemplo de una pregunta del cuestionario.

El banco de preguntas estaba compuesto por 660 preguntas, una para cada una de las imágenes del *dataset* (ver Tabla 2). El cuestionario que se presentó al participante contenía 42 preguntas: 6 preguntas para cada una de las 7 emociones (6 preguntas × 7 emociones= 42 preguntas) que fueron seleccionadas aleatoriamente del banco de preguntas. El orden de las preguntas presentadas a los participantes fue aleatorio, al igual que el orden en que se presentaban las opciones de respuesta de cada una de las preguntas.

Tabla 2: Número de preguntas por emoción del banco de preguntas

Emoción	Numero de preguntas
Neutral	20
Ira	80
Asco	60
Miedo	80
Alegría	280
Tristeza	120
Sorpresa	20

3.3 Procedimiento

Después de explicar los objetivos del experimento, se instruyó a los participantes para que se sentaran y respondieran el cuestionario con las cuarenta y dos preguntas. Se les indicó que observaran la imagen presentada en la pregunta e intentaran identificar la emoción correspondiente a la expresión facial del avatar sintético que aparecía en la pregunta.

La respuesta al cuestionario tuvo una duración de unos 10 minutos por participante.

3.4 Resultados

En esta sección presentamos los resultados de clasificación obtenidos en términos de precisión y matriz de confusión. La precisión es una medida de rendimiento global calculada como la proporción de clasificaciones correctas sobre el total de clasificaciones realizadas. Una matriz de confusión es una tabla utilizada para describir el desempeño de una clasificación que resume el número de clasificaciones correctas e incorrectas para cada una de las clases.

La clasificación de emociones realizada por los participantes obtuvo una precisión global de 0.56.

La Figura 3 muestra la matriz de confusión obtenida con la clasificación de emociones realizada por los participantes humanos. Los mejores resultados de clasificación fueron para las emociones **neutral** (88.2%) y **sorpresa** (84.8%), mientras que la peor emoción identificada fue la emoción de **miedo** (25.8%).
Las emociones más habitualmente confundidas fueron miedo con sorpresa (36.4%) y miedo con neutral (22%).

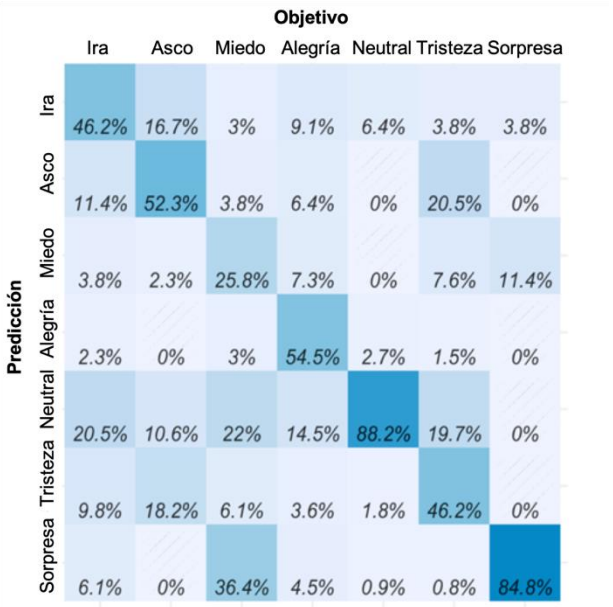

Figura 2: Matriz de confusión de la clasificación de las emociones realizada por los participantes humanos.

4. El experimento con modelos de aprendizaje automático

En esta sección detallamos el experimento realizado para analizar cómo un modelo de aprendizaje automático reconoce las emociones a partir de imágenes de avatares sintéticos realizando expresiones faciales.

4.1 El modelo CNN

En este experimento, usamos una red neuronal convolucional (CNN) simple con tres capas de un patrón de convolución + *max-pooling*, seguido de dos capas densas o completamente

conectadas. La entrada a la red es una imagen en escala de grises de 128x128, con un *stride* de S=1 y *zero-padding* en las capas de convolución, asegurándonos así de que las dimensiones espaciales de la entrada no se alterarán y dejando la tarea del submuestreo a la capa de *pool*. En la capa de *pool*, usamos un *max-pooling* de 2x2 para reducir las dimensiones de la entrada. Finalmente, el último tensor de salida de la base convolucional se aplana en una capa densa. La capa densa final tiene las 7 salidas correspondientes a nuestras 7 clases de emociones. El esquema de la CNN se resume en la Figura 4.

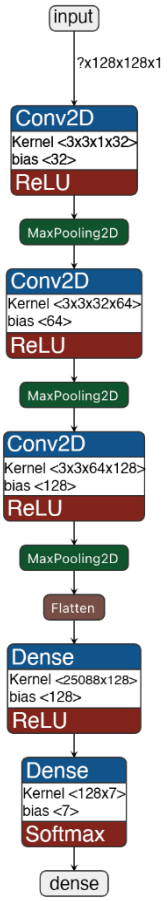

Figura 4: El modelo de red neuronal convolucional.

4.2 Procedimiento

Los pasos de pre-procesamiento incluyen el recorte de la cara, para reducir el posible efecto del fondo de la imagen sobre el entrenamiento de la red neuronal, la conversión de la imagen a escala de grises y el cambio de tamaño de la imagen para que se ajuste al tamaño de 128x128 píxeles de los datos de entrada de la CNN.

Tabla 3: Número de imágenes por emoción en los datasets: dataset completo, dataset de training y dataset de test

Emoción	UIBVFED	Training	Test
Neutral	20	16	4
Ira	80	64	16
Asco	60	48	12
Miedo	80	64	16
Alegría	280	224	56
Tristeza	120	96	24
Sorpresa	20	16	4

Para la fase de entrenamiento del modelo CNN usamos el 80% de los datos incluidos en el dataset y para la fase de test usamos el 20% restante de los datos (ver Tabla 3).

4.3 Resultados

El modelo CNN obtuvo una precisión general de 0.88 y los resultados de la clasificación se resumen en la matriz de confusión que se muestra en la Figura 5.

Los mejores resultados de clasificación fueron para las emociones de **alegría** (100%), **ira** (93.8%) y **miedo** (87.5%), mientras que la peor emoción identificada fue la **neutral** (0%). Las emociones más habitualmente confundidas fueron neutral con tristeza (75%) y sorpresa con miedo (50%).

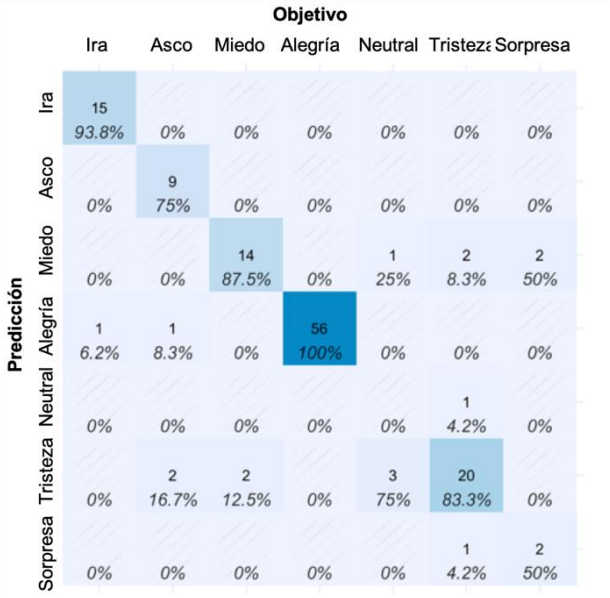

Figura 5: Matriz de confusión de la clasificación de emociones realizada por el modelo CNN.

5. Discusión

El rendimiento de la clasificación en términos de precisión mostrado por los participantes humanos podría considerarse un rendimiento bajo (0.56). Se obtuvieron buenos resultados (superiores a 0.8) en el reconocimiento de emociones como la neutral y la sorpresa, aunque resultados muy pobres a la hora de reconocer expresiones como la de miedo (sobre 0.25).
Además de la información de cuáles son las emociones mejores y peor clasificadas, también es interesante la información de cuáles son las emociones más comúnmente confundidas. Para intentar encontrar una explicación a estas confusiones entre las emociones, en la Tabla 4 hemos relacionado las similitudes entre las diferentes expresiones faciales a partir de las características faciales de cada una de las expresiones de los estudios de Faigin (Faigin, 2012) y, de ellas, hemos extrapolado las similitudes entre emociones. Las conclusiones a las que hemos llegado pueden consultarse de forma resumida en la Tabla 5.

En el caso de las emociones clasificadas por participantes humanos, las emociones más habitualmente confundidas fueron la de miedo con sorpresa (36.4%), un resultado que ya fue previsto por Faigin (ver Tabla 4 y Tabla 5), y cuya explicación radica en que las expresiones faciales de ambas emociones comparten las características de boca abierta y cejas levantadas.

El segundo par de emociones más habitualmente confundido fue la de miedo con neutral (22%). A pesar de que esta confusión no está explícitamente contemplada en el análisis de expresiones faciales similares que hace Faigin, debemos fijarnos en que nuestros participantes también confunden la emoción de tristeza con la neutral en un 19.7% de casos (ver líneas continuas en la Figura 6). Esta transitividad entre la confusión de las emociones miedo-neutral-tristeza parece indicar una confusión indirecta entre las emociones de miedo y tristeza (ver líneas discontinuas en la Figura 6), la cual si que aparece indicada en los estudios de Faigin. Esta confusión indirecta podría explicarse por la similitud entre las emociones miedo-neutral y tristeza-neutral en los niveles menos intensos de sus expresiones faciales asociadas, donde los músculos faciales se encuentran más relajados y, por lo tanto, más cercanos a la expresión facial neutral (ver Figura 7).

Si miramos detalladamente la Figura 7, donde aparecen las expresiones faciales menos intensas asociadas a las emociones miedo, neutral y tristeza, podemos ver que la única diferencia observable a simple vista en las tres expresiones faciales (miedo, neutral y triste) es un ligero cierre gradual de los ojos, así como un sutil cambio en la posición de las cejas; diferencias difícilmente distinguibles para un observador humano.

Tabla 4: Similitud entre las expresiones faciales y sus emociones asociadas (ver Tabla 1)

Emoción	Expresión facial	Expresión facial similar	Emoción similar
Ira	Ira con labios comprimidos	Tristeza reprimida	Tristeza
	Ira severa	-	-
	Ira gritando	-	-
	Enojo	-	-
Asco	Desdén	-	-
	Asco	-	-
	Repulsión	-	-
Miedo	Miedo	Triste	Tristeza
	Terror	Sorpresa	Sorpresa
	Muy asustado	Sorpresa	Sorpresa
	Preocupado	Tristeza reprimida	Tristeza
Alegría	Risa falsa 1	Risa	Alegría
	Sonrisa falsa	Sonrisa con la boca abierta	Alegría
	Sonrisa con la boca cerrada	Sonrisa falsa	Alegría
	Sonrisa con la boca abierta	Sonrisa falsa	Alegría
	Sonrisa sofocada	-	-
	Risa	Llorando con la boca abierta Risa falsa 2	Tristeza Alegría
	Risa estruendosa	-	-
	Risa falsa 2	Risa	Alegría
	Sonrisa avergonzada	-	-
	Sonrisa ansiosa	Sonrisa falsa	Alegría
	Sonrisa congraciadora	-	-
	Sonrisa astuta	-	-
	Sonrisa melancólica	Triste	Tristeza
	Sonrisa desleal	-	-
Tristeza	Llorando con la boca cerrada	Sonrisa sofocada	Alegría
	Llorando con la boca abierta	Risa	Alegría
	Miserable	Preocupado	Miedo
	Casi llorando	-	-
	Triste	Preocupado Neutral	Miedo Neutral
	Tristeza reprimida	-	-
Sorpresa	Sorpresa	Miedo	Miedo

Tabla 5: Similitud entre emociones (resumen de la Tabla 4).

Emoción	Emoción similar
Neutral	-
Ira	Tristeza
Asco	-
Miedo	Tristeza Sorpresa
Alegría	Alegría Tristeza
Tristeza	Alegría Miedo Neutral
Sorpresa	Miedo

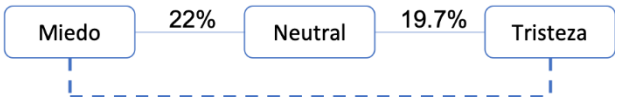

Figura 6: Relación entre las confusiones de las emociones miedo, neutral y tristeza.

Miedo Neutral Triste

Figura 7: Ejemplos de imágenes del mismo avatar realizando la expresión facial de miedo (emoción: miedo), neutral (emoción: neutral) y triste (emoción: tristeza).

En la Figura 8 pueden observarse, ordenadas ascendentemente por grado de intensidad, todas las expresiones faciales asociadas a las emociones de miedo, neutral y tristeza.

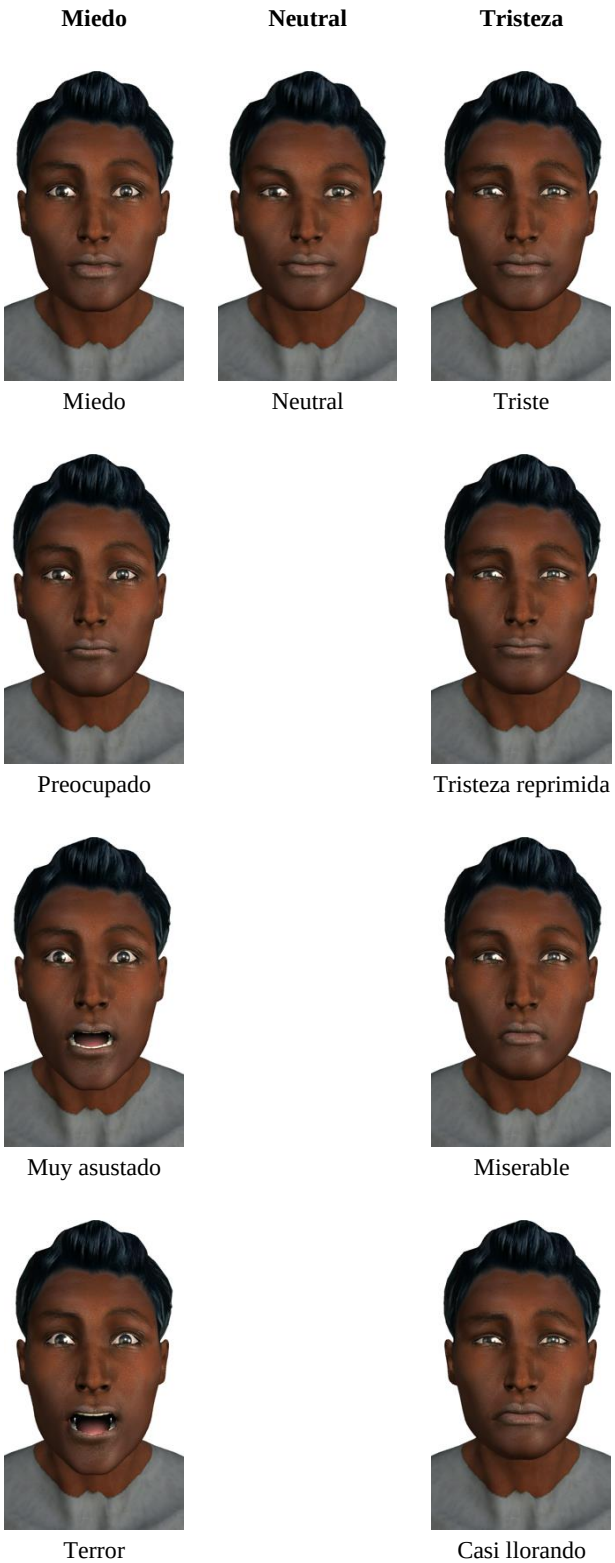

Figura 8: Ejemplos de imágenes del mismo avatar realizando todas las expresiones faciales asociadas a las emociones de miedo (primera columna), neutral (columna central) y tristeza (tercera columna), ordenadas por grado de intensidad (de menor a mayor).

Miedo	Neutral	Tristeza
		 Llorando con la boca cerrada
		 Llorando con la boca abierta

Figura 8 (Cont.): Ejemplos de imágenes del mismo avatar realizando todas las expresiones faciales asociadas a las emociones de miedo (primera columna), neutral (columna central) y tristeza (tercera columna), ordenadas por grado de intensidad (de menor a mayor).

Con el modelo de la CNN se obtuvo un rendimiento global de 0.88, un 57% superior al obtenido con participantes humanos. En este caso, los mejores porcentajes de reconocimiento (superiores al 90%) se obtuvieron para las emociones de alegría e ira y los peores resultados fueron para la emoción neutral. Es destacable que no coincidan ninguna de las emociones (ni las mejores ni la peor reconocida) con el comportamiento mostrado por los participantes humanos. Respecto a las emociones más habitualmente confundidas por la CNN: neutral-tristeza (75%) y sorpresa-miedo (50%), ambas confusiones son esperadas según el análisis de Faigin (ver Tabla 4 y Tabla 5) y, en este caso, sí que detectamos una coincidencia entre el comportamiento de la CNN y el comportamiento de los participantes humanos con la confusión entre sorpresa y miedo.

A la hora de entender el porqué de las confusiones de la CNN al clasificar las emociones, nos encontramos con uno de los mayores retos a la hora de analizar las predicciones proporcionadas por una red neuronal, su explicabilidad. La interpretabilidad de los resultados obtenidos es difícil debido a los robustos y complejos algoritmos que hay detrás del funcionamiento de una red neuronal. Habitualmente, nos vemos obligados a confiar en el resultado obtenido y a obviar el proceso que ha llevado hasta él. Esta necesidad de explicación llevó a la aparición de las técnicas de XAI (Explainable Artificial Intelligence), las cuales permiten al usuario obtener información sobre el proceso que ha seguido un modelo para determinar la respuesta que se ha obtenido (Adadi & Berrada, 2018).

Una de las técnicas de XAI más utilizadas con las CNN, y que permite una explicación sobre la imagen de entrada, es el sistema LIME (Ribeiro, Singh, & Guestrin, 2016). LIME es una abreviación de Local Interpretable Model-Agnostic Explanations. Es un modelo local porque la explicación que obtenemos se refiere a un caso concreto para el que se ha obtenido una predicción y es un modelo agnóstico porque puede explicar cualquier decisión sin necesidad de conocer el funcionamiento interno del modelo que se analiza. Para determinar las partes de la entrada que contribuyen a la predicción, LIME perturba ligeramente la entrada a la red neuronal y observa cómo se comporta la nueva predicción. Después, los puntos que se han perturbado se ponderan de acuerdo con su proximidad a la muestra original y se utilizan para el aprendizaje de un modelo lineal a partir del cual se interpretará la predicción. Para perturbar la imagen original, LIME la divide en un conjunto de super-píxeles o regiones (ver la Figura 9), y determina cuáles de ellos participan en la justificación de la elección de una predicción.

Figura 9: Ejemplo de una imagen mostrando las áreas de la imagen divididas por super-píxeles (nº de super-píxeles = 50).

Por ejemplo, si consideramos la imagen de la Figura 10, correspondiente a la expresión facial asociada a la emoción de sorpresa y que nuestra CNN predice incorrectamente como la emoción de miedo, y queremos determinar qué zonas de la imagen han influido en la respuesta proporcionada por el modelo, podemos hacerlo a través de LIME. La imagen de la

Figura 11 muestra la explicación proporcionada por LIME resaltando en azul los super-píxeles o áreas de la imagen que han tenido peso sobre la predicción. Observamos que las áreas de la imagen que llevan a la predicción se centran en la parte de los ojos, la boca y la frente, áreas que coinciden con las zonas donde se sitúan las características faciales distintivas de las diferentes expresiones faciales. En este caso concreto, la predicción de miedo se produce al observar en la imagen las características de boca abierta, ojos abiertos y cejas levantadas, las cuales forman parte de algunas de las expresiones faciales asociadas a la emoción de miedo (además de a la expresión facial de sorpresa). Por lo tanto, esta confusión concreta de la CNN entre miedo-sorpresa es coherente con las emociones similares de Faigin.

Figura 10: Avatar generando la expresión facial asociada a la emoción de sorpresa.

Figura 11: Explicación proporcionada por LIME de una predicción de una imagen correspondiente a la emoción de sorpresa (expresión facial: sorpresa) incorrectamente etiquetada como miedo.

Como ejemplo de explicación para una expresión correctamente clasificada, la Figura 12 resalta las zonas en las que se ha fijado la CNN para clasificar correctamente la emoción de miedo a partir de la expresión facial "muy asustado". Se ve la influencia que tienen en la predicción las zonas de la boca y los ojos. En concreto: boca abierta, ojos muy abiertos y cejas levantadas y juntas.

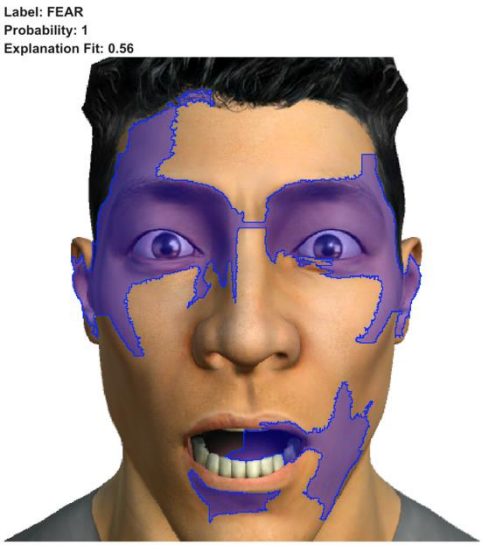

Figura 12: Explicación proporcionada por LIME de una predicción de una imagen correspondiente a la emoción de miedo (expresión facial: muy asustado) correctamente etiquetada como miedo.

La Figura 13 corresponde a la explicación de la clasificación incorrecta de una emoción neutral como una emoción de tristeza. Como ya habíamos comentado, la emoción neutral, si nos fijamos en la zona de los ojos y la boca, es fácilmente confundible con el nivel menos intenso de la expresión facial asociada a la emoción de tristeza (triste), muy parecida a la neutral.

Figura13: Explicación proporcionada por LIME de una predicción de una imagen correspondiente a la emoción neutral (expresión facial: neutral) incorrectamente etiquetada como tristeza.

Figura 14: Explicación proporcionada por LIME de una predicción de una imagen correspondiente a la emoción de tristeza (expresión facial: triste) correctamente etiquetada como tristeza.

La Figura 14 corresponde a la explicación de una predicción correcta de la expresión facial triste como emoción de tristeza. La red neuronal ha llegado a esta explicación fijándose únicamente en la parte de los ojos, ligeramente cerrados por la presión hacia abajo del pliegue oblicuo con un movimiento hacia arriba del párpado inferior, y las cejas. Tiene sentido que no considere la parte de la boca, ya que esta expresión facial se caracteriza por una boca relajada.

Debemos tener en cuenta que, en todos los ejemplos, las explicaciones generadas son a nivel de super-píxel, es decir, como puede verse en la Figura 9, un super-píxel puede no corresponderse exactamente con una característica facial, es decir, un ojo o la boca en su totalidad. La Figura 14 muestra claramente cómo parte de la zona del ojo izquierdo del personaje incluye también una parte de la frente, esto es debido a que algunos de los super-píxeles que incluyen el ojo contienen estas zonas adicionales (ver Figura 9).

Ante los resultados obtenidos, podemos deducir que las CNN realizan un análisis más detallado de la imagen de entrada que el análisis visual que realiza un humano al observar una imagen, pudiendo fijarse más en las pequeñas sutilezas (detectadas en forma de características por los distintos filtros que aplica nuestra red neuronal) y, por lo tanto, pudiendo distinguir mejor entre expresiones faciales similares. Por lo que el modelo CNN, aunque sea simple, al poder fijarse más en los detalles de la imagen, se asemeja más al comportamiento teórico esperado según Faigin.

6. Conclusión

El reconocimiento de expresiones faciales es un tema frecuentemente abordado desde el surgimiento de la Inteligencia Artificial y, generalmente, los principales esfuerzos suelen estar dirigidos a aumentar la precisión de clasificación de los modelos de aprendizaje automático entrenados. Pero, si tenemos en cuenta que el objetivo original de la Inteligencia Artificial no era crear una inteligencia perfecta sino crear una inteligencia que se comportara como los humanos, nos preguntamos si, además de centrarnos en las medidas de rendimiento de los modelos de aprendizaje automático para validar un modelo, también deberíamos centrarnos en lo parecido que es su comportamiento al comportamiento humano.

Realizamos un experimento con participantes humanos para analizar cómo los humanos reconocen las emociones al observar una expresión facial en un avatar sintético estático, obteniendo una precisión general de 0.56, lo que se consideraría como un mal resultado de rendimiento para un modelo de aprendizaje automático.

Los resultados de este experimento nos dan una idea de lo difícil que es, incluso para los humanos, distinguir correctamente las emociones a partir de imágenes estáticas sin contexto. Como ya indicó Faigin (Faigin, 2012), existen expresiones faciales que tienen acciones faciales similares, lo que explica que existan emociones fácilmente confundibles entre ellas, como pueden ser el miedo y la sorpresa. En el caso de nuestros participantes humanos, las emociones más

habitualmente confundidas fueron miedo con sorpresa y miedo con neutral.

Una vez analizado cómo los humanos reconocen las emociones a partir de expresiones faciales estáticas, entrenamos un modelo simple de CNN consiguiendo una precisión global de 0.88, un valor 57% superior a la precisión obtenida con participantes humanos. Pero, aunque este valor se consideraría un buen resultado desde el punto de vista del rendimiento global del modelo, debemos notar que las emociones mejor y peor clasificadas difieren de las emociones mejor y peor clasificadas por los participantes humanos, así como la mayoría de las confusiones entre emociones detectadas. Por lo tanto, podemos concluir que el comportamiento de clasificación mostrado por nuestro modelo CNN es diferente al comportamiento de clasificación mostrado por nuestros participantes humanos.

En concreto, nuestro modelo CNN puede realizar un análisis más detallado de las imágenes y fijarse en características más sutiles a la hora de discernir entre las diferentes expresiones, por lo que puede obtener tasas de reconocimiento superiores a las que hemos detectado entre los participantes humanos y un comportamiento más cercano al comportamiento teórico predicho por Faigin. Por lo tanto, hemos llegado a un modelo cuyo comportamiento puede considerarse más correcto teniendo en cuenta las bases teóricas de reconocimiento de emociones a partir de expresiones faciales, pero es diferente al comportamiento mostrado por nuestros participantes humanos.

Intentando responder a la pregunta inicial de "¿Deberíamos conseguir un clasificador de expresiones faciales infalible o deberíamos conseguir un clasificador de expresiones faciales que fuera indistinguible de un humano, es decir, que fuera capaz de superar el *Test de Turing*?", queda probado que, en el caso del reconocimiento de expresiones faciales, el comportamiento de clasificación humano dista mucho del comportamiento de clasificación teórico. Por lo tanto, es difícil conseguir un modelo de clasificación de expresiones faciales que sea "perfecto" desde el punto de vista teórico y, al mismo tiempo, muestre el mismo patrón de reconocimiento que nuestros participantes humanos. La elección entre intentar conseguir el comportamiento teórico o el comportamiento humano dependerá del objetivo al que el modelo sea destinado: ¿queremos crear una inteligencia "perfecta" o queremos crear una inteligencia que imite a la humana? Recordemos que ya en 1995 Russell & Norvig contemplaban ambos enfoques en su definición universal de lo que es la Inteligencia Artificial (Russell & Norvig, 1995).

7. Limitaciones y trabajo futuro

Aunque la discusión acerca de intentar crear clasificadores con un alto grado de precisión y la conveniencia (o no) de analizar su similitud con el comportamiento humano podría aplicarse en todos aquellos campos donde se intente sustituir o replicar el comportamiento humano, este estudio se ha realizado en el caso concreto de expresiones faciales realizadas por avatares sintéticos. Por lo tanto, sería interesante analizar si el mismo comportamiento observado (tanto del modelo como de los participantes humanos) se reproduce en el caso de la clasificación de expresiones faciales realizadas por humanos.

Agradecimientos

Los autores agradecen el proyecto EXPLainable Artificial INtelligence systems for health and well-beING (EXPLAINING) financiado por PID2019-104829RA-I00 / MCIN/ AEI / 10.13039/501100011033. También agradecemos su apoyo a la Universidad de las Islas Baleares y al Departamento de Ciencias Matemáticas e Informática.

Referencias

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Bettadapura, V. (2012). Face expression recognition and analysis: the state of the art. *ArXiv Preprint ArXiv:1203.6722*.

Chirra, V. R. R., Uyyala, S. R., & Kolli, V. K. K. (2021). Virtual facial expression recognition using deep CNN with ensemble learning. *Journal of Ambient Intelligence and Humanized Computing*, *12*(12), 10581–10599.

Colbois, L., de Freitas Pereira, T., & Marcel, S. (2021). On the use of automatically generated synthetic image datasets for benchmarking face recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1–8).

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80. https://doi.org/10.1109/79.911197

del Aguila, J., González-Gualda, L. M., Játiva, M. A., Fernández-Sotos, P., Fernández-Caballero, A., & García, A. S. (2021). How Interpersonal Distance Between Avatar and Human Influences Facial Affect Recognition in Immersive Virtual Reality. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.675515

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology \& Nonverbal Behavior*.

Faigin, G. (2012). *The artist's complete guide to facial expression*. Watson-Guptill.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). 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* (pp. 94–101).

Lundqvist, D., Flykt, A., & Öhman, A. (1998). Karolinska directed emotional faces. *Cognition and Emotion*.

Lyons, M. J., Kamachi, M., & Gyoba, J. (1997). Japanese female facial expressions (JAFFE). *Database of Digital Images*, *3*.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31.

Oliver, M. M., & Amengual Alcover, E. (2020). UIBVFED: Virtual facial expression dataset. *PLOS ONE*, *15*(4), 1–10. https://doi.org/10.1371/journal.pone.0231266

Ramis, S., Buades, J. M., Perales, F. J., & Manresa-Yee, C. (2022). A Novel Approach to Cross dataset studies in Facial Expression Recognition. *Multimedia Tools and Applications*, 1–38.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR*, *abs/1602.0*. Retrieved from http://arxiv.org/abs/1602.04938

Russell, S., & Norvig, P. (1995). Artificial intelligence: A Modern appraoch. Prentice-Hall.

Skjuve, M., Haugstveit, I. M., Følstad, A., & Brandtzaeg, P. (2019). Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human–chatbot interaction. *Human Technology*, *15*(1), 30–54. https://doi.org/https://doi.org/10.17011/ht/urn.201902201607

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Valstar, M., Pantic, M., & others. (2010). Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect* (p. 65).